Contribution à l'analyse de la dynamique des écritures anciennes pour l'aide à l'expertise paléographique

Hani Daher

Université Paris Descartes
UFR de Mathématiques et Informatique
ED Informatique, Télécommunication et Electronique

Thèse présentée pour l'obtention du grade de
Docteur de l'université Paris Descartes
Discipline : **Informatique**

Sujet de Thèse :

Contribution à l'analyse de la dynamique des écritures anciennes pour l'aide à l'expertise paléographique

---

Présentée par :
Hani Daher

Directeur de thèse : **Pr. Nicole Vincent**
Co-encadrement de thèse : **Dr. Véronique Eglin**

Soutenue le 22 novembre, devant le jury composé de :

Pr. Rolf Ingold	Université de Fribourg	Rapporteur
Pr. Jean-Marc Ogier	Université de la Rochelle	Rapporteur
Dr. Laurence Likforman-Sulem	TélécomParisTech	Examinateur
Pr. Thierry Paquet	Université de Rouen	Examinateur
Pr. Nicole Vincent	Université Paris Descartes	Directrice de thèse
Dr. Véronique Eglin	Institue Nationale des Sciences Appliquées-Lyon	Co-encadrente de thèse

REMERCIEMENTS

RÉSUMÉ DE THÈSE

Mes travaux de thèse s'inscrivent dans le cadre du projet ANR GRAPHEM[1] (*Grapheme based Retrieval and Analysis for PaleograpHic Expertise of Middle Age Manuscripts*). Ils présentent une contribution méthodologique applicable à l'analyse automatique des écritures anciennes pour assister les experts en paléographie dans le délicat travail d'étude et de déchiffrage des écritures.

L'objectif principal est de contribuer à une instrumetation du corpus des manuscrits médiévaux détenus par l'Institut de Recherche en Histoire des Textes (IRHT – Paris) en aidant les paléographes spécialisés dans ce domaine dans leur travail de compréhension de l'évolution des formes de l'écriture par la mise en place de méthodes efficaces d'accès au contenu des manuscrits reposant sur une analyse fine des formes décrites sous la formes de petits fragments (les graphèmes). Dans mes travaux de doctorats, j'ai choisi d'étudier la dynamique de l'élément le plus basique de l'écriture appelé le ductus[2] et qui d'après les paléographes apporte beaucoup d'informations sur le style d'écriture et l'époque d'élaboration du manuscrit.

Mes contributions majeures se situent à deux niveaux : une première étape de prétraitement des images fortement dégradées assurant une décomposition optimale des formes en graphèmes contenant l'information du ductus. Pour cette étape de décomposition des manuscrits, nous avons procédé à la mise en place d'une méthodologie complète de suivi de traits à partir de l'extraction d'un squelette obtenu à partir de procédures de rehaussement de contraste et de diffusion de gradients. Le suivi complet du tracé a été obtenu à partir de l'application des règles fondamentales d'exécution des traits d'écriture, enseignées aux copistes du Moyen Age. Il s'agit d'information de dynamique de formation des traits portant essentiellement sur des indications de directions privilégiées.

Dans une seconde étape, nous avons cherché à caractériser ces graphèmes par des descripteurs de formes visuelles compréhensibles à la fois par les paléographes et les informaticiens et garantissant une représentation la plus complète possible de l'écriture d'un point de vue géométrique et morphologique. A partir de cette caractérisation, nous avons proposé une approche de clustering assurant un regroupement des graphèmes en classes homogènes par l'utilisation d'un algorithme de classification non-supervisé basée sur la coloration de graphe. Le résultat du clustering des graphèmes a conduit à la formation de dictionnaires de formes caractérisant de manière individuelle et discriminante chaque manuscrit traité. Nous avons également étudié la puissance discriminatoire de ces descripteurs afin

---

[1] Projet ANR, 2007-2011, http://liris.cnrs.fr/graphem/

[2] Ductus : ordre et sens du tracé de chaque partie de la lettre, élément fondamental pour l'étude de l'évolution des écritures

d’obtenir la meilleure représentation d’un manuscrit en dictionnaire de formes. Cette étude a été faite en exploitant les algorithmes génétiques par leur capacité à produire de bonne sélection de caractéristiques.

L’ensemble de ces contributions a été testé à partir d’une application CBIR sur trois bases de manuscrits dont deux médiévales (manuscrits de la base d’Oxford et manuscrits de l’IRHT, base principale du projet), et une base comprenant de manuscrits contemporains utilisée lors de la compétition d’identification de scripteurs d’ICDAR 2011. L’exploitation de notre méthode de description et de classification a été faite sur une base contemporaine afin de positionner notre contribution par rapport aux autres travaux relevant du domaine de l’identification d’écritures et étudier son pouvoir de généralisation à d’autres types de documents. Les résultats très encourageants que nous avons obtenus sur les bases médiévales et la base contemporaine, ont montré la robustesse de notre approche aux variations de formes et de styles et son caractère résolument généralisable à tout type de documents écrits.

Mots clés : Paléographie, ductus, dynamique de l’écriture, diffusion du gradient, suivi du tracé, coloration de graphe, sélection de caractéristiques par algorithmes génétiques, dictionnaires de formes, CBIR, binarisation, rehaussement de contraste, segmentation.

ABSTRACT

My thesis work is part of the ANR GRAPHEM Project (*Grapheme based Retrieval and Analysis for Expertise paleographic Manuscripts of Middle Age*). It represents a methodological contribution applicable to the automatic analysis of ancient writings to assist the experts in paleography in the delicate work of the studying and deciphering the writing.

The main objective is to contribute to an instrumentation of the corpus of medieval manuscripts held by "*I*nstitut de *R*echerche en *H*istoire de *T*extes" (IRHT-Paris), by helping the paleographers specialized in this field in their work of understanding the evolution of forms in the writing, with the establishment of effective methods to access the contents of manuscripts based on a fine analysis of the forms described in the form of small fragments (graphemes). In my PhD work, I chose to study the dynamic of the most basic element of the writing called the ductus and which according to the paleographers, brings a lot of information on the style of writing and the era of the elaboration of the manuscript.

My major contribution is situated at two levels: a first step of preprocessing of severely degraded images to ensure an optimal decomposition of the forms into graphemes containing the ductus information. For this decomposition step of manuscripts, we have proceeded to the establishment of a complete methodology for the tracings of strokes by the extraction of the skeleton obtained from the contrast enhancement and the diffusion of the gradient procedures. The complete tracking of the strokes was obtained from the application of fundamental execution rules of the strokes taught to the scribes of the Middle Ages. It is related to the dynamic information of the formation of strokes focusing essentially on indications of the privileged directions.

In a second step, we have tried to characterize the graphemes by visual shape descriptors understandable by both the computer scientists and the paleographers and thus unsuring the most complete possible representation of the wrting from a geometrical and morphological point of view. From this characterization, we have have proposed a clustering approach insuring a grouping of graphemes into homogeneous classes by using a non-supervised classification algorithm based on the graph coloring. The result of the clustering of graphemes led to the formation of a codebook characterizing in an individual and discriminating way each processed manuscript. We have also studied the discriminating power of the descriptors in order to obtain a better representation of a manuscript into a codebook. This study was done by exploiting the genetic algorithms by their ability to produce a good feature selection.

The set of the contributions was tested from a CBIR application on three databases of manuscripts including two medieval databases (manuscripts from the Oxford and IRHT databases), and database of containing contemporary manuscripts used in the writers

identification contest of ICDAR 2011. The exploitation of our description and classification method was applied on a cotemporary database in order to position our contribution with respect to other relevant works in the writrings identification domain and study its generalization power to other types of manuscripts. The very encouraging results that we obtained on the medieval and contemporary databases, showed the robustness of our approach to the variations of the shapes and styles and its resolutely generalized character to all types of handwritten documents.

Keywords: paleography, ductus, writing dynamic, gradient diffusion, stroke tracking, graph coloring, feature selection by genetic algorithms, codebooks, CBIR, binarization, contrast enhancemet, segmentation.

TABLE DES MATIERES

Introduction 1
1 Contexte des travaux de thèse 1
2 Organisation du mémoire 5
3 L'écriture d'un point de vue paléographique 6
3.1 Caractéristiques paléographiques des traits 6
3.2 Définition de l'ordre des traits et de la densité des encres dans le Ductus 7
3.3 Étude de l'épaisseur et de la décomposition des traits 8
3.4 Étude de la dynamique du tracé pour la décomposition des traits 8
3.5 Le poser et lever du calame 10
4 Descriptions des bases utilisées 10
4.1 La base de manuscrits de l'IRHT 11
4.2 Base de manuscrits d'Oxford 13
4.3 Base de manuscrits contemporains 14
Chapitre 1 : État de l'art sur les méthodes de classification et de reconnaissance des styles d'écritures 16
1 Introduction 16
1.1 Les scénarios d'analyse dans le domaine de l'analyse des écritures 16
1.2 Les étapes fondamentales des systèmes de classification et de reconnaissance de styles 18
2 Taxonomie de description de l'information manuscrite pour la discrimination de styles et des scripteurs 21
2.1 Description par connexités : d'une description locale à l'élaboration d'une signature globale de la page 22
2.2 Méthodes de discrimination de styles basées sur une description des mots ou des lignes 26
2.3 Méthodes de discrimination de styles basées sur une description texture des textes. 28
2.4 Bilan sur les méthodes d'identification de manuscrits 31
3 Les algorithmes de classification supervisés et non-supervisés (Catégorisation de manuscrits en styles) 32
3.1 Les méthodes supervisées 33
3.2 Les méthodes de classification non-supervisées 43
3.3 Méthodes de classification semi-supervisées 53
3.4 Bilan sur les méthodes de classification supervisées, non-supervisées et semi-supervisées 54
4 Choix de la méthode de coloration de graphe 55
4.1 Flexibilité de la méthode de coloration de graphe 57
4.2 Comparaison de la méthode de coloration de graphe avec les autres méthodes 58
5 Conclusion 60
Chapitre 2 : État de l'art sur les méthodes de caractérisation des écritures globales, locales et mixtes 62
1 Introduction 62
2 Les approches globales 65
2.1 L'analyse de texture 65
2.2 L'analyse fréquentielle 67
2.3 L'analyse multi-résolution et multi-échelle 71
2.4. Analyse fractale 74
2.5 Analyse par la loi de Zipf 76

3 Approches locales et description statistique 78
3.1 Analyse par les contours 78
3.2 Analyse par approximation polygonale 81
3.3 Analyse basée sur le squelette 82
4 Approches mixtes et analyse statistique des décompositions locales des écritures 85
5 Conclusion 88
Chapitre 3 : Une Approche structurelle pour la construction de dictionnaires de formes 91
1 Introduction 91
2 État de l'art 92
2.1 Prétraitement 92
2.2 Méthodes de squelettisation 106
3 Suivi du tracé et analyse de l'axe médian 109
3.1 Principe général de notre approche de suivi du tracé 110
4 Décomposition en traits 118
4.1 Étude de la stabilité de la décomposition 122
5 Conclusion 125
Chapitre 4 : Caractérisation des graphèmes et construction de dictionnaire de formes 128
1 Introduction 128
2 Travaux antérieurs 129
2.1 Les familles de méthodes de sélection de caractéristiques 129
2.2 Sélection et pondération des caractéristiques 133
2.4 Principe général des algorithmes génétiques utilisés pour la sélection de caractéristiques 134
3 Construction d'un dictionnaire de formes associé à un manuscrit 136
3.1 Principe général de l'approche de construction de dictionnaire de formes par AG 136
3.2 Choix des descripteurs initiaux 138
3.3 Sélection ou pondération de caractéristiques par AG 153
4 Principe théorique de la coloration de graphes 158
4.1 Modélisation du problème de classification de graphèmes en termes de coloration 158
4.2 Construction du dictionnaire de formes 159
5 Comparaison de la sélection et pondération 160
6 Résultats et application 163
6.1 De la sélection de caractéristiques à la déduction des poids génériques 164
6.2 Validation des poids génériques pour toutes les classes 165
6.3 Poids génériques et dictionnaire de formes associés à un ensemble de documents 168
7 Conclusion 170
Chapitre 5 : Exploitation des dictionnaires de formes pour la comparaison des écritures 172
1 Introduction 172
2 Mesures de similarité 173
2.1 Un peu d'histoire 173
2.2 Les images : un stimulus particulier 177
3 Dictionnaires de formes et méthodes de comparaison 177
3.1 Représentation théorique des dictionnaires de formes 177
3.2 Distance de Hausdorff 178
3.3 Distance de Hausdorff modifiée (Schaefer) 179
3.4 Distance perpétuelle modifiée de Hausdorff 179
3.5 Earth Mover's Distance 179
3.6 Distance de corrélation pondérée 180
3.7 Distance Quadratique 181

4 Application à la reconnaissance de styles, basée sur la technique CBIR 181
4.1 Comparatifs des performances des trois principales mesures de similarités entre dictionnaires : DHD, DHDM et DQ 182
4.2 Analyse des performances de notre contribution 184
5 Comparaison des résultats portant sur les poids génériques et les dictionnaires de formes représentatifs de chaque style 186
6 Positionnement de notre méthode : compétition ICDAR 2011 187
6.1 Description des méthodes 187
6.2 Notre contribution en deux méthodes 188
6.3 Description de la base ICDAR 2011 189
6.4 Métrique d'évaluation 189
6.5 Évaluation des méthodes 190
7 Conclusion 195
Conclusion 198
1 Conclusion générale 198
2 Perspectives 199
2.1 Les perspectives à courts termes 200
2.2 Les perspectives à moyen et long termes 201
Publications 204
Bibliographie 205

LISTE DES FIGURES

Figure I.1. Exemples des manuscrits anciens latins du Moyen Age, IRHT 1
Figure I.2. (a) Vieillissement de l'encre et papier, (b) Enchevêtrement des lignes, (c) Ecriture dans la marge et/ou entre les lignes 3
Figure I.3. Ductus de la lettre "W" 7
Figure I.4. Variation de l'épaisseur du trait selon l'orientation et la direction du trait (rouge pointillé) 8
Figure I.5. Différence, dynamique et forme 9
Figure I.6. Représentation dynamique du tracé 9
Figure I.7. (a)Poser et lever de calame qui suivent la règle d'exécution («haut-gauche, bas-droite»), (b) cas de poser et lever qui s'écartent de la règle générale 10
Figure I.8. Classes et nombre de manuscrits constituant la base de manuscrits de l'IRHT 11
Figure I.9. Echantillons de manuscrits représentant les 22 classes de la base réduite de l'IRHT 12
Figure I.10. Exemples des 4 classes de manuscrits de la base d'Oxford 13
Figure I.11. Exemple d'un texte produit en quatre langues : (a) anglais, (b) grec, (c) français, (d) allemand, (ICDAR 2011) 14
Figure 1.1. Illustration des différents scénarios de classification et de reconnaissance de scripteurs et de styles d'écritures *(Atanasiu et al. [2011])* 18
Figure 1.2. Les étapes fondamentales d'un système de reconnaissance de styles d'écriture .. 19
Figure 1.3. Variabilité du « a » de la base de manuscrits de l'IRHT 20
Figure 1.4. Représentation des deux familles de classification supervisée et non-supervisée. 33
Figure 1.5. Hyperplan à marge-maximale et des marges pour un *SVM* formé avec des échantillons de deux classes. Les échantillons sur la marge sont appelés les vecteurs supports 36
Figure 1.6. Modèle graphique du perceptron multicouche *(Dawson et Wilby, [2001])* 41
Figure 1.7. Principe de la méthode des nuées dynamiques avec ($k \leq 2$), *(Diday, [1971])* 46
Figure 1.8. Schéma d'une carte auto-organisatrice *(Kohonen, [1982])* 52
Figure 1.9. Exploitation de la coloration de graphe (a) pour la construction des dictionnaires de formes, (b) pour la classification des manuscrits à partir d'un clustering des dictionnaires de formes 58
Figure 1.10. Evaluation de la classification *(Gaceb et al. [2009])* 59
Figure 1.11. Comparaison des trois classifieurs *(Gaceb et al. [2009])* 59
Figure 2.1. Résumé des caractéristiques extraites selon les mécanismes globaux, locaux et mixtes considérant le document à différents niveaux d'intérêt et d'échelle 64
Figure 2.2. Rendu visuel de *4* styles d'écritures paléographiques (IRHT) 65
Figure 2.3. (a) Orientation du gradient, (b) courbure gaussienne 67
Figure 2.4. Exemple de la transformée de Fourier d'un document manuscrit 68
Figure 2.5. Exemples de filtrages par la transformée de Fourier et des fréquences sélectionnées *(Dargenton, [1991])* 68
Figure 2.6. (a) Exemple de signature d'un extrait de manuscrit en 6 directions principales, (b) signature de *10* scripteurs sur des extraits du corpus de Montesquieu (qui comporte plus de 30), *(Eglin et al. [2006])* 70
Figure 2.7. Évolution multi-résolution de la cursivité d'une portion de texte *(Eglin et al. [2004])* 72
Figure 2.8. Représentation des courbures et orientations de la forme "*R*", *(Joutel et al. [2008])* 73
Figure 2.9. Signature de l'écriture proposée par *(Joutel et al. [2008])* 73
Figure 2.10. Apparence d'un graphe d'évolution *(Boulétreau et al. [1998])* 75

Figure 2.11. Graphe de lisibilité *(Boulétreau, [1998])* 76
Figure 2.12. Courbe de Zipf associée à un manuscrit avec extraction des zones linéaires conduisant au calcul des paramètres de la loi, caractéristiques de l'écriture, *(Pareti et Vincent, [2006])* 77
Figure 2.13. (a) code de Freeman avec *8* directions (*8*-connexité) ; (b) code de Freeman avec *4* directions (*4*-connexité)..... 81
Figure 2.14. Représentation du contour à partir de:(a) code de Freeman, (b) Polygone 82
Figure 2.15. Exemples de dictionnaires de formes. (a) (Bensefia et al. [2005b]), (b) (Bulacu et Schomacker, [2005]) 86
Figure 2.16. Découpage du mot headlines à l'aide de fenêtres glissantes 87
Figure 3.1. Image Originale extraite de la base de manuscrits médiévaux de l'IRHT 94
Figure 3.2. Résultats de binarisation avec un seuil global de binarisation: (a) Isodata *(Velasco, [1980]),* (b) Entropie Max *(Cheng et al. [1998]),* (c) Otsu *(Otsu, [1979]),* (d) Entropie minimum *(Li et Lee, [1993]),* (e) Moments *(Tsai, [1985]),* (f) Moyenne *(Glasbey, [1993])* 95
Figure 3.3. Méthodes de binarisation locales appliquées au document de la figure 1 96
Figure 3.4. Résultats de binarisation par la méthode d'Otsu à la suite de l'application de différentes techniques de rehaussement de contraste. L'axe des abscisses présente les *8* images de la base..... 104
Figure 3.5. Résultats des méthodes de rehaussement sur l'image 7 de la base DIBCO11 105
Figure 3.6. Schéma général de principe d'extraction de l'axe médian : du rehaussement au découpage..... 111
Figure 3.7. (a) image originale, (b) algorithme feu de brousse avec $l = 1$ et $\theta\left(\nabla I^{n=8}\right)$, (c) $l = 40$ et $\theta\left(\nabla I^{n=82}\right)$ 115
Figure 3.8. Régularisation et diffusion du gradient sur un extrait de texte ancien..... 115
Figure 3.9. Extraction de l'axe médian par, (a) notre méthode, (c) la méthode de Zhang. (b) binarisation par la méthode de Sauvola, (d) résultat de notre méthode sur un manuscrit dégradé (encre claire, bruit, trait effacé) 116
Figure 3.10. Illustration de l'algorithme de suivi du tracé et extraction de l'axe médian 118
Figure 3.11. Cette figure montre d'une part (a) les règles de formations des traits et d'autre part, (b) les règles de décomposition..... 119
Figure 3.12. Points de décomposition aux points minimaux locaux des traits 119
Figure 3.13. (a) Exemple de décomposition des traits en graphèmes par notre méthode, (b) Courbes des épaisseurs des points de l'axe médian des lettres « O » et « C», et points de découpage en graphèmes, (c) résultats de découpage sur d'autres styles d'écritures 120
Figure 3.14. (a) ordre des traits, (b) point de décomposition indiquant une variation dans l'intensité du gradient, (c) exemples de décomposition de deux lettres « *X* » aux points de croisement produisant trois segments 121
Figure 3.15. Illustration des points de jonction simple. (a) ordre de visite des traits, (b) exemple d'intersection sur des manuscrits médiévaux 122
Figure 3.16. Exemples des lettres les plus utilisées dans les manuscrits 122
Figure 3.17. Schéma de formation des occurrences de la lettre « a » sur huit manuscrits..... 123
Figure 3.18. Exemple d'occurrences de la lettre « a » et leur segmentation pour le manuscrit 1 avec $C = 3$ et $S_{M1} = 0,5$ 123
Figure 3.19. Valeurs de *S* pour les lettres « *a, b, d, e, f, o* », L'axe des abscisses présente les 8 manuscrits retenus pour mener les tests 124
Figure 3.20. Valeurs de la stabilité *S* pour les lettres « a, b, d, e, f, o ». Pour toutes les lettres, le meilleur taux de décomposition est donné par le premier cas (cas 1)..... 125
Figure 4.1. Procédure générale de sélection de caractéristiques 130
Figure 4.2. Les catégories de méthodes de sélection de caractéristiques..... 130

Figure 4.3. Modèle général de sélection de caractéristiques à partir des méthodes de filtrage. Les caractéristiques sont filtrées indépendamment de l'algorithme d'induction 131
Figure 4.4. Modèle général de sélection de caractéristiques à partir des méthodes enveloppantes 131
Figure 4.5. Illustration des opérateurs de croisement et mutation *(Huang et Wang, [2006])* 135
Figure 4.6. (a) Caractérisation et sélection de caractéristiques par algorithme génétique combiné à la coloration de graphes, (b) déduction des poids et test de validité du système 138
Figure 4.7. (De gauche à droite), Hauteur et largeur d'un graphème, excentricité, densité globale, direction, périmètre 139
Figure 4.8. Evolution des graphèmes en fonction de la hauteur et de la largeur 142
Figure 4.9. Evolution de l'excentricité des graphèmes en fonction de transformations 143
Figure 4.11. Evolution de la direction des graphèmes en fonction de transformations 145
Figure 4.12. Evolution du périmètre des graphèmes en fonction de transformations 146
Figure 4.13. Evolution de la circularité et de la compacité des graphèmes en fonction de transformations 147
Figure 4.14. Evolution des 9 densités des graphèmes en fonction de transformations 148
Figure 4.15. Calcul de D_{10} pour *5* graphèmes, les résultats sont donnés (a) avant normalisation, (b) après normalisation 149
Figure 4.16. Evolution des graphèmes en fonction des *9* orientations, avec les cartes de chaleurs représentant les *9* directions pour chacun des 5 graphèmes 150
Figure 4.17. Evolution des graphèmes en fonction des *25* Moments de Zernike 151
Figure 4.18. (a) Codage des paramètres de sélection en deux parties séparées d'un chromosome, (b) Codage des paramètres de pondération en deux parties séparées d'un chromosome 154
Figure 4.19. Principes de croisement et de mutation entre deux chromosomes 157
Figure 4.20. Etapes de construction du dictionnaire de formes par coloration de graphe 159
Figure 4.21. Courbes représentant les fitness maximales sur les 20 images en utilisant les techniques expliquées dans les sections précédentes 161
Figure 4.22. Evolution de la fonction de fitness et du seuil de coloration 162
Figure 4.23. (a)Relation entre poids et caractéristiques, (b) poids et descripteurs 162
Figure 4.24. Les quatre styles présents dans la base d'Oxford 163
Figure 4.25. Courbes représentant les valeurs de la fonction de fitness qui sont calculées à partir des poids génériques et à partir de la sélection de caractéristiques sur les *n* (*n=100*) manuscrits de la base d'apprentissage 164
Figure 4.26. Résultats des valeurs de fitness à l'issue des 3 tests 165
Figure 4.27. Relation entre poids et caractéristiques 166
Figure 4.28. Résultat de construction des dictionnaires de formes utilisant la coloration de graphes et la pondération de caractéristiques sur des manuscrits de la base Oxford et IRHT 167
Figure 4.29. (a) distribution des poids pour toutes les classes de la base B1, (b) distribution des poids pour chaque classe de la base B1 169
Figure 5.1. Exemple de non respect de l'inégalité triangulaire. L'image de gauche *(A)* et l'image de droite *(C)* sont jugées relativement dissemblables. En revanche, celle du milieu *(B)* est à la fois similaire à *(A)* et à *(C).* La distance *d(A,C)* serait donc supérieure à la somme *d(A,B) + d(B,C).* 176
Figure 5.2. (a) Manuscit *m*, (b) représentation du dictionnaire de formes 178
Figure 5.3. Courbes de précision-rappel, base Oxford 183
Figure 5.4. Courbes de précision-rappel, base IRHT 185

Figure 5.5. Résultats des poids génériques et dictionnaires de formes spécifiques à chaque style et poids génériques sur toute la base sans dictionnaires de formes représentatifs. 186
Figure 5.6. Quatre échantillons de la base ICDAR 2011 écrits en : (a) anglais, (b) français, (c) allemand, (d) grec 189
Figure C.1. Décomposition, (a) schéma graphique, (b) partitions musicales, les traits en couleurs présentent les graphèmes extraient à partir de ces images 199
Figure C.2. Différentes décompositions de la lettre « a » dans un même manuscrit (IRHT) 202

LISTE DES TABLEAUX

Tableau 1.1. Résultats du classifieur par *kppv* pour la reconnaissance de styles d'écriture 35
Tableau 1.2. Résultats des *SVM* sur la reconnaissance de styles d'écritures 37
Tableau 1.3. Résultats des Arbres de décision sur la classification des manuscrits 38
Tableau 1.4. Résultats des Réseaux bayésiens sur la classification des manuscrits 40
Tableau 1.5. Résultats des Réseaux de neurones sur la classification des manuscrits 43
Tableau 1.6. Résultats des méthodes incrémentatles pour l'identification de scripteurs 45
Tableau 1.7. Résultats des méthodes *k*-moyennes, *k*-médoïdes et nuées dynamiques sur la classification des manuscrits 47
Tableau 1.8. Résultats des méthodes de classification hiérarchique sur la classification des manuscrits 50
Tableau 1.9. Résultats des cartes auto-organisatrices sur la classification des manuscrits 53
Tableau 1.10. Résultats des méthodes semi-supervisées pour la classification de manuscrits et l'identification de scripteurs 54
Tableau 2.1. Résumé des caractéristiques globales utilisées dans le domaine de l'identification de scripteurs et de la classification de manuscrits 78
Tableau 2.2. Résumé des caractéristiques locales utilisées dans le domaine de l'identification de scripteurs et de la classification de manuscrits 84
Tableau 2.3. Résumé des approches mixtes utilisées dans le domaine d'identification de scripteurs et de classification de manuscrits 88
Tableau 3.1. Résumé des méthodes de binarisation : globales, locales et hybrides 99
Tableau 3.2. Résumé des méthodes de rehaussement de contraste avec : G(Global), L(Local), LN (Linéaire), NL (Non linéaire), S(Spatial), F(Fréquentiel) 103
Tableau 3.3. Avantages et inconvénients des 6 méthodes de squelettisation 108
Tableau 3.4. Bilan sur les méthodes de squelettisation 109
Tableau 3.5. Modèles possibles en *2D* pour les valeurs propres λ_1 et λ_2 de la matrice Hessienne avec *P* = valeur petite, *E* = valeur élevée, +/- indiquent le signe des valeurs propres selon la condition $|\lambda_1| \leq |\lambda_2|$ 112
Tableau 4.1. Résumé des méthodes de sélection de caractéristiques 132
Tableau 4.3. Invariance des descripteurs à l'échelle et rotation 152
Tableau 4.4. Distances entre les poids des classes 169
Tableau 5.1. Résultats de précision et rappel pour les trois distances sur les bases de manuscrits d'Oxford et de l'IRHT 183
Tableau 5.2. Résultats de précision et rappel pour les trois distances sur les bases de manuscrits d'Oxford et de l'IRHT 184
Tableau 5.3. Evaluation souple en utilisant toute la base (%) 191
Tableau 5.4. Evaluation stricte en utilisant toute la base (%) 191
Tableau 5.5. Evaluation souple sur les manuscrits grecs (%) 191
Tableau 5.6. Evaluation souple sur les manuscrits anglais(%) 191

Tableau 5.7. Evaluation souple sur les manuscrits français(%) 192
Tableau 5.8. Evaluation souple sur les manuscrits allemands(%) 192
Tableau 5.9. Evaluation souple en utilisant toute la base 192
des images recadrées (%) 192
Tableau 5.10. Evaluation stricte en utilisant toute la base 193
des images recadrées(%) 193
Tableau 5.11. Evaluation souple sur les 193
manuscrits grecs recadrés (%) 193
Tableau 5.12. Evaluation souple sur les 193
manuscrits anglais recadrés (%) 193
Tableau 5.13. Evaluation souple sur seulement les 194
manuscrits français recadrés (%) 194
Tableau 5.14. Evaluation souple sur les 194
manuscrits allemands recadrés(%) 194
Tableau 5.15. Classement général en fonction des scores de *S* pour toutes les expériences.
Les colonnes *2* à *13* reprennent les résultats des tableaux *2* à *13* précédents 194

Introduction

1 Contexte des travaux de thèse

Cette thèse s'inscrit dans le cadre du projet ANR GRAPHEM (*Grapheme based Retrieval and Analysis for PaleograpHic Expertise of Middle Age manuscripts*). Elle représente une contribution méthodologique applicable à l'analyse automatique des écritures manuscrites anciennes pour assister les experts en paléographie dans le délicat travail d'étude et de déchiffrage des écritures. Nous nous sommes intéressé aux manuscrits latins du Moyen Age qui précèdent la période de la Renaissance, avant l'émergence de l'imprimerie (Figure I.1).

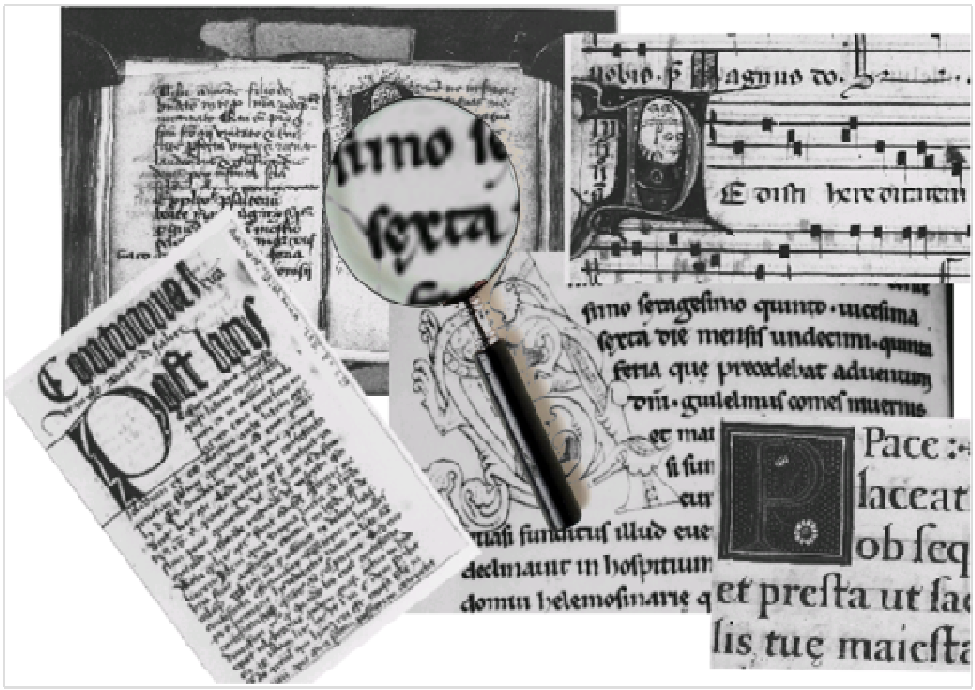


Figure I.1. Exemples des manuscrits anciens latins du Moyen Age, IRHT

La production de l'écrit, au Moyen Age, présente la particularité d'avoir des procédures raisonnées de fabrication et de contrôle de qualité. L'analyse nécessite de chercher les indices les plus subtiles présents dans les manuscrits, et de faire « parler » les signes visuels présents dans les formes et les mots. En ce sens, il faut bien comprendre la production de l'écrit du Moyen Age comme un art à la fois visuel et manuel basé sur la réplication d'un ensemble de signes et de formes. L'enquête paléographique et philologique révèle, à travers l'histoire de la

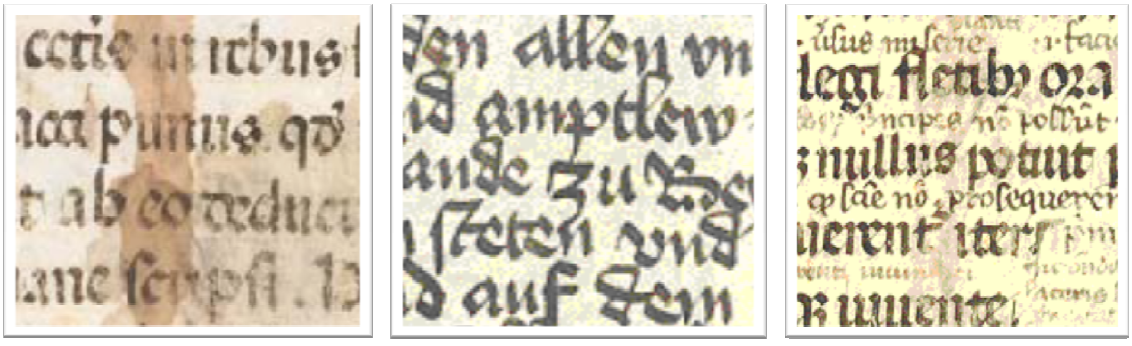
production écrite, que les textes du Moyen Age sont autant de formes ou de signes que de contenus relatant la culture ancienne. L'étude de l'évolution de l'écriture a conduit les paléographes à s'intéresser de près aux spécialisations des ateliers de copies et à comprendre les circuits des influences graphologiques qui sont aussi celles des idées et des savoirs.

Le projet GRAPHEM s'inscrit dans ce contexte d'études savantes de l'évolution des écritures et va chercher à mettre au point des méthodes et des techniques qui permettront d'aider les experts paléographes à retrouver et comprendre les origines souvent énigmatiques des manuscrits anciens. Dans ce travail, nous ne nous intéressons pas spécifiquement au sens dégagé par les textes, mais à l'analyse des traits d'écriture qui sont perçus ici comme des marques permettant d'en définir la provenance, d'aider les experts à resituer le texte dans son contexte historique ou géographique et d'en exploiter le contenu. L'analyse du texte et l'exploitation de son contenu portent ainsi sur l'analyse de l'écriture et de ses particularités morphologiques. La notion de graphème est donc centrale dans cette thèse. Elle est à mettre en relation directe avec l'idée d'une décomposition de l'écriture en petits fragments ou en petites unités qui constituent l'écriture dans son ensemble et participent à l'impression générale qui s'en dégage.

Sur le million environ de manuscrits médiévaux conservés dans le monde, très peu sont à ce jour décrits avec la précision souhaitable, moins de *10%* dans les bibliothèques majeures. Un immense travail reste donc à réaliser pour éditer les textes en étudiant notamment les nombreuses sources inédites, de manière à mettre leur contenu à la disposition de la communauté scientifique. Une très grande partie de ces documents du Moyen Age sont détenus et numérisés par l'Institut de Recherche et d'Histoire des Textes (IRHT), partenaire de ce projet. Plusieurs centaines de milliers de manuscrits médiévaux numérisés en mode image attendent d'être analysés pour être rendus accessibles sur Internet à des chercheurs du monde entier. Contrairement aux documents imprimés numérisés dont les contenus textuels peuvent être automatiquement reconnus par OCR (Optical Character Recognition), les documents médiévaux totalement manuscrits ou peints sont actuellement hors de portée des processus de traitements automatisés. Cette description des contenus des images est nécessaire pour le fonctionnement des moteurs de recherche sur Internet (Google, Yahoo, Quaero…) et des bibliothèques numériques (Bibliothèque Nationale de France « Gallica », Bibliothèque Numérique Européenne « BNUE », Europeana). Cependant, il existe un trop petit nombre d'experts face à une trop grande quantité de documents à analyser pour que le minutieux travail de description puisse être réalisé dans des délais raisonnables. Les outils informatiques sont alors nécessaires non seulement pour assister les historiens dans leur travail de recherche mais aussi pour augmenter l'efficacité et la précision de leurs méthodes de travail. Les objectifs de cette étude sont multiples et interviennent à plusieurs niveaux d'analyse. Les principaux obstacles qu'il faut

lever dans ce travail sont liés au manque crucial d'outils informatiques pour l'assistance du travail de description et l'analyse experte dans les disciplines des Sciences Humaines. Le développement de traitements informatisés destinés aux experts pour les assister dans leur travail de description des contenus ne peut se faire que sur la base d'une collaboration étroite entre les chercheurs en Sciences Humaines et Sociales (SHS) et les chercheurs en Informatique (STIC). Pour les chercheurs en Sciences Humaines il ne s'agit pas de remettre en cause leurs méthodes de travail mais d'intégrer une dimension systémique dans l'analyse des formes et des ductus dont ils étaient jusqu'ici privés afin de permettre à leurs expertises d'être plus objectives qu'elles ne le sont actuellement. La collaboration avec les chercheurs en Informatique a notamment nécessité pour les chercheurs en SHS de décrire leurs méthodes actuelles de travail et de les ouvrir aux traitements informatisés des images.

Grâce à l'assistance de l'analyse d'images, l'extraction de descripteurs de formes et aux techniques de classification, le passage d'une étude ponctuelle sur quelques ouvrages à une recherche sur une plus grande échelle (plusieurs milliers d'ouvrages) a pu être envisagé pour les chercheurs en SHS offrant ainsi l'espoir de découvertes rendues impossibles à l'échelle de l'ouvrage. Sur ce type de manuscrits, on observe des particularités spécifiques aux documents anciens, comme le vieillissement des supports et des encres, l'imprégnation irrégulière des encres, les plissements, déchirements, cassures et autres dégradations du papier (figure I.2). De plus, les règles d'exécution des écritures en paléographie sont très strictes : certaines lettres et combinaisons de lettres ne peuvent être produites que selon une unique dynamique d'exécution. Celle-ci est parfaitement maîtrisée par les paléographes qui connaissent très précisément les contraintes de formation des lettres et de construction des ductus. Il est donc nécessaire de tenir compte dans notre étude de toutes ces contraintes et ces particularités d'exécution des écritures.



(a) (b) (c)

Figure I.2. (a) Vieillissement de l'encre et papier, (b) Enchevêtrement des lignes, (c) Ecriture dans la marge et/ou entre les lignes

Dans le domaine de l'analyse des écritures, on distingue différents objectifs. On peut s'intéresser à l'identification d'un individu, d'un copiste que l'on nomme le scripteur, en

s'intéressant aux échantillons d'écritures présentant de fortes similarités de contenus. Il s'agit dans ce cas de porter l'attention sur des particularités internes spécifiques d'une main. L'identification du scripteur cherche à tirer profit de la variabilité des autres écritures avec lesquelles il faut produire la comparaison. Le domaine de l'authentification des mains qui fait partie de ce domaine et a ainsi toute sa place dans les mécanismes de datation des écrits. De manière complémentaire aux tâches d'identification, on peut également s'intéresser aux approches de vérification du scripteur. Alors que l'identification d'un scripteur consiste à identifier un individu parmi un ensemble de scripteurs connus du système, la tâche de vérification consiste à déterminer si deux échantillons d'écriture sont ou non le produit d'une même main. Dans ce travail de thèse, nous nous sommes intéressé à l'analyse des écritures au sens du style (à la différence du scripteur qui est perçu comme une empreinte de la personnalité individuelle) dans un contexte de corpus d'images. Nous nous intéressons ici à l'appartenance d'une écriture à une famille présentant des propriétés morphologiques communes et qui se basent sur l'estimation de critères décrivant ce qui, dans l'écriture, est invariant (stable et fréquent) plutôt que spécifique et rare. La notion de style d'écriture est donc centrale dans notre travail. Le style d'une écriture est caractérisé par la présence de formes redondantes distribuées sur la surface totale de l'échantillon d'écriture, et qu'il est possible de classifier selon des critères perceptuels graphométriques spécifiques (rondeur, cursivité, linéarité…) rendant compte de leur fréquence d'apparition. Deux textes de styles différents peuvent ainsi être décrits avec le même vocabulaire de base contenant l'ensemble des occurrences des formes élémentaires (sacs de mots) mais conduisant à un rendu très différent en fonction à la fois des combinaisons locales des mots de ce vocabulaire et de leur fréquence d'apparitions sur la page.

Le style est une notion très perceptive associée à l'apparence de l'écriture dans sa globalité. Deux échantillons d'écriture de styles similaires seront donc perceptuellement proches et devront pouvoir être associés par le classifieur dans une même classe de style. Dans ce contexte, l'objectif du travail de thèse consiste à proposer des classements paramétrables des écritures selon leur style. La notion de style paléographique est une notion qui a été laissée volontairement floue par les experts paléographes associés à cette étude afin de ne pas influencer le système automatique d'extraction de descripteurs et le fonctionnement du classifieur. Il s'agit pour nous de trouver des réponses possibles à la question : « Existe-t-il une description robuste des écritures, capable de conduire à une classification des styles d'écritures paléographiques sans connaissance a priori sur les spécificités grapho-morphologiques des écritures à travers les époques ? » Cette question soulève naturellement la question de la pertinence des descripteurs, de l'élaboration de mesures de similarités consistantes et de la mise en place d'un outil de classification capable de produire des jeux de classes discriminants. Les enjeux de la thèse touchent à ces trois étapes fondamentales que tout classifieur met en jeu.

Parmi les applications de l'analyse des styles des écritures paléographiques, nous avons attaché une attention particulière à la recherche d'information par le contenu (RI) dans les corpus d'images de textes, à la classification des styles en groupes homogènes et à l'identification d'un style au sein d'un ensemble de styles connus a priori. Ces différents aspects applicatifs seront présentés dans le dernier chapitre de la thèse à travers l'analyse de plusieurs bases d'images de manuscrits de la période médiévale. Nous démontrerons enfin la généralisation de notre méthode de classification en styles sur des images de textes manuscrits contemporains.

Concrètement il s'agit pour nous :

- De produire une décomposition de l'écriture en graphèmes cohérents, en évitant notamment de produire des graphèmes qui correspondraient à certains gestes de rebroussement (retour en arrière du mouvement de la plume), qui sont considérés comme des mouvements incompatibles avec la nature des plumes (le plus souvent des calames) et celle des supports (nature du papier) à ces époques.
- De produire une classification de l'ensemble des graphèmes produisant un dictionnaire de formes (nommé également codebook), considéré comme un dictionnaire des graphèmes triés par similarité. Cette classification est destinée à un usage paléographique : elle intègre la possibilité pour les experts en Sciences Humaines de proposer simplement plusieurs solutions de classification des écritures par la saisie d'une seule valeur de seuil et offre la possibilité d'un rendu visuel exploitable par les experts par reprojection des étiquettes (ou couleurs) des graphèmes sur les pages d'écritures. Les similarités rendues ainsi visibles facilitent la saisie des fragments de lettres fréquents et des formes redondantes.
- De classifier les manuscrits par style d'écriture ou par main en se basant sur l'exploitation des dictionnaires de formes, dictionnaires considérés comme des signatures propres à chaque manuscrit.

2 Organisation du mémoire

Le mémoire s'articule en cinq chapitres.

Le premier chapitre présente un état de l'art sur les méthodes d'identification de styles d'écriture existantes dans la littérature. Il est divisé en deux parties. Dans la première partie nous présentons d'une manière générale les différentes méthodes d'identification d'écriture et les familles auxquelles elles appartiennent. Dans la seconde partie nous présentons les algorithmes de classification qui sont au cœur de chacune de ces approches. Nous discuterons en dernière partie de notre choix d'algorithme de classification. Ce choix est fait pour atteindre les objectifs de bonne discrimination des écritures et de formation de classes cohérentes.

Le deuxième chapitre décrit les différents types de caractéristiques (globales, locales et hybrides) que l'on peut extraire sur des images d'écritures et plus généralement sur des images de traits. Pour chaque famille de caractéristiques nous présentons une description précisant la façon dont elle est calculée, une interprétation du sens de la caractéristique (en terme de rendu visuel : rondeur, courbure, linéarité, cursivité…) et le rôle de cette caractéristique dans le domaine de l'identification d'écritures.

Le troisième chapitre, présente notre méthode de décomposition des écritures en graphèmes (définis comme les traits élémentaires de l'écriture) passant par des étapes de prétraitement d'images inspirées d'approches essentiellement utilisées en imagerie médicale et en analyse d'images naturelles : pour le rehaussement du contraste local en bordure des formes et l'estimation puis le suivi de l'axe médian des traits d'écritures. L'étape de décomposition du manuscrit en graphèmes est quant à elle inspirée de mécanismes issus de l'analyse de la dynamique de formation du tracé en suivant des critères de décomposition spécifiques à la formation des écrits médiévaux et appris des paléographes.

Le quatrième chapitre contient plusieurs parties : la caractérisation des graphèmes, la sélection et la pondération des caractéristiques. Dans ce chapitre nous décrivons chacune des caractéristiques utilisées pour caractériser les graphèmes, présentons aussi le concept de dictionnaire de formes utilisé comme signature des manuscrits. A partir de la caractérisation des graphèmes et de l'utilisation des dictionnaires de formes nous montrons comment notre approche offre à la fois une vue locale et globale d'un manuscrit.

Le cinquième et dernier chapitre, permet de situer notre travail dans son contexte applicatif. Nous présentons les résultats de recherche d'information selon des mécanismes CBIR sur les différentes bases. Nous montrerons en quoi notre méthode peut être adaptée à n'importe quel type de manuscrits et qu'elle n'est pas seulement dédiée aux manuscrits médiévaux. Des perspectives abordant les aspects d'identification d'écritures et de classification selon les mêmes mécanismes de description, de partitionnement en classes et d'apprentissage seront également présentés.

3 L'écriture d'un point de vue paléographique

3.1 Caractéristiques paléographiques des traits

La production du document manuscrit du Moyen Age dépend de plusieurs facteurs : le choix du support (parchemin ou papier), l'utilisation de constituants matériels spécifiques (encres et couleurs), la formation du ductus (formation des lettres). Ainsi, le trait laissé par un calame, une plume d'oiseau ou une plume métallique produira un rendu tout différent. Nous présentons ici quelques termes utilisés en paléographie et dont nous ferons usage dans le manuscrit de thèse :

1. **Paléographie :** la paléographie s'attache à déterminer les rapports existant entre divers dialectes, la façon dont ils dérivent les uns des autres et les règles que suivent les évolutions phonétiques, grammaticales et sémantiques.

2. **Le ductus :** le ductus représente le nombre, l'ordre et la direction des traits pendant la formation d'une lettre. On va considérer l'exemple de la lettre « w » (figure I.3). Cette lettre est formée de 4 traits, de gauche à droite (ordre) et de bas en haut (direction). On pourra ainsi s'apercevoir, pourvu que le trait soit assez long pour rendre visible le phénomène, que l'intensité du coloris diminue au fur et à mesure que l'encre s'épuise. Cela constitue un indicateur important aidant la détermination du sens de l'écriture du trait considéré, on parle alors de "ductus des traits". L'analyse du ductus constitue un domaine d'expertise spécifique de paléographie.

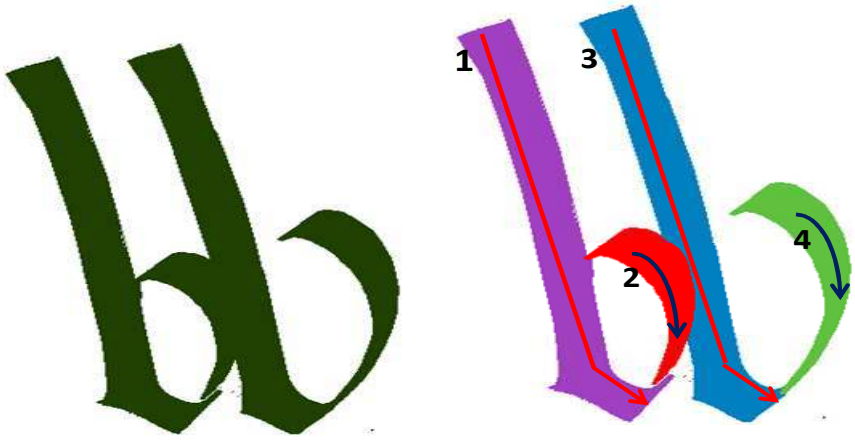


Figure I.3. Ductus de la lettre "W"

3. **Le calame :** le calame représente une plume d'oiseau ou un roseau effilé, taillé avec un couteau. Il est considéré comme l'objet privilégié de réalisation de l'écriture durant le Moyen Age.

3.2 Définition de l'ordre des traits et de la densité des encres dans le Ductus

En général, l'exécution d'un trait manuscrit se fait de gauche à droite : le premier trait est situé à gauche et puis les autres traits le suivent selon la direction gauche-droite.

La densité de l'encre est également utilisée pour retrouver l'ordre du ductus à condition de pouvoir en percevoir les différences. L'indication d'une goutte sombre ou claire dans un trait peut aider à retrouver l'ordre du tracé. Le problème dans ce genre d'interprétation vient essentiellement de la nature et de la qualité du support : s'il est dégradé, il sera d'autant plus difficile d'observer ces différences. On peut aussi utiliser la densité de l'encre pour déterminer les posers et levers de calames. Au début du trait, la densité d'encre est importante (cas du

calame sans réservoir d'encre) puis au cours du tracé la densité devient moins importante jusqu'à la fin du trait où on rencontre la densité d'encre la plus élevée (comme un effet « goutte » de fin de tracé).

3.3 Étude de l'épaisseur et de la décomposition des traits

L'observation de l'écriture montre que l'épaisseur du trait varie selon la direction de la ligne d'après la règle illustrée dans la figure 4. Cette règle est utilisée par les paléographes de l'IRHT durant l'analyse des manuscrits.

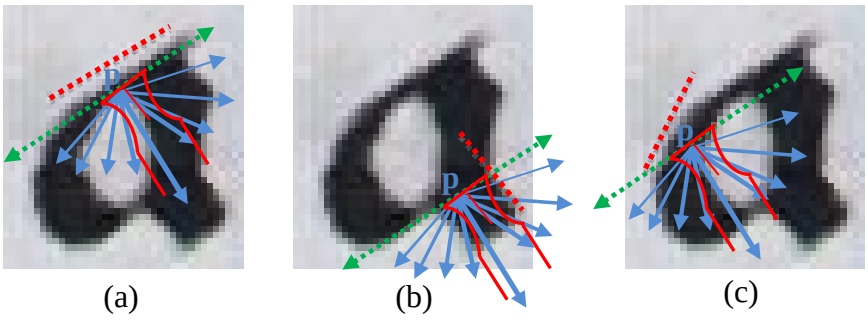


Figure I.4. Variation de l'épaisseur du trait selon l'orientation et la direction du trait (rouge pointillé)

La figure I.4 contient la règle générale régissant la formation des épaisseurs des traits selon différentes directions. Tous les traits "noirs" ayant une direction comprise entre *[-135°, 60°]*, et s'effectuant de gauche à droite et de haut en bas obéissent à la règle générale. Le trait supérieur (à environ *60°*) pointillé représente la frontière avec une zone de tracé à l'intérieur de laquelle les traits s'épaississent au fur et à mesure du changement de direction (figure I.4 (a)). Le trait le plus épais se trouve dans l'axe ayant une direction perpendiculaire à l'axe pointillé (figure 4 (b)). Au fur et à mesure des changements d'orientation des traits, on voit les épaisseurs s'affiner à nouveau jusqu'à l'arrivée au trait qui représente la frontière dans le même axe pointillé que précédemment (figure I.4 (c)). En exploitant au mieux ce principe d'évolution de l'épaisseur des traits, il devient possible de déterminer leur direction. Par ailleurs, il est difficile de connaître la position de la plume au cours de l'écriture du ductus, mais il est possible de connaître la position du bout de la plume sur le support. De manière générale, il est impossible de dire si l'angle d'inclinaison de la plume a une influence sur l'épaisseur du trait.

3.4 Étude de la dynamique du tracé pour la décomposition des traits

Il est nécessaire de faire la différence entre la forme et la dynamique dans l'exécution de la forme. Une simplification toute arbitraire du problème pourrait ainsi être contenue dans la règle de base qui dit qu' « il y a une différence entre ce qu'on voit (le rendu) et ce qu'on fait (la manière de produire le trait) ». Par exemple, dans la figure suivante, la forme du *"S"* qui est à

gauche ne nous permet pas de connaître ni le nombre, ni l'ordre, ni la direction des traits comme on les voit dans le *"S"* de droite (figure I.5).

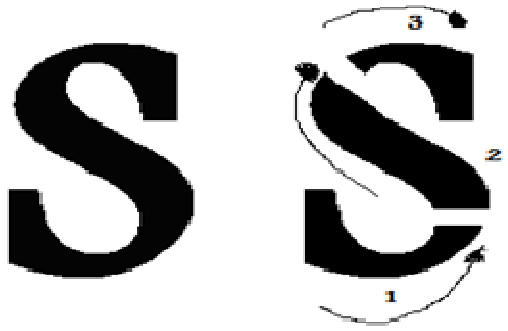


Figure I.5. Différence, dynamique et forme

La figure I.6 illustre une représentation possible des aspects dynamiques du tracé en rapport avec la façon dont la forme a été exécutée. Nous avons représenté sur cette image les levers et des posers du calame dont la position est notée par les cercles rouges. Considérons maintenant la zone qui est dans le carré rouge. Si on commence par exemple à examiner le tracé à partir du point L, lorsqu'on arrive au point Pk, on observe un changement de direction. Ces indices nous aident à analyser le sens de l'exécution du trait. Dans cet exemple, le parcours n'a pas pu se faire dans la direction (L-P_k) avec un calame et cela nous conduit à considérer la direction opposée. De même la présence de forte courbure le long de l'axe central nous indique la présence éventuelle d'un poser ou d'un lever de calame (cercle vert). Dans tous les cas, il est nécessaire d'apprendre le geste, de savoir le reproduire et le contrôler pour l'interpréter. L'expertise paléographique est essentielle ici.

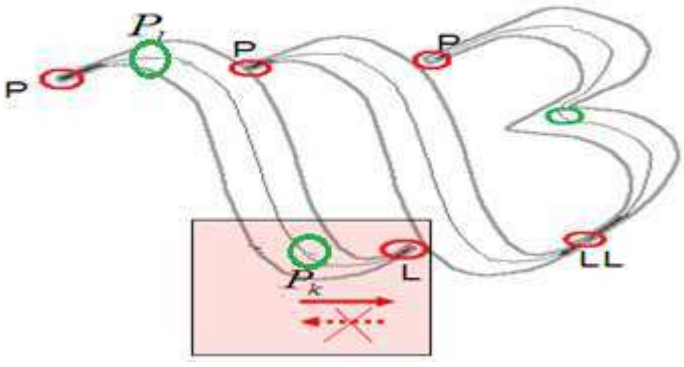


Figure I.6. Représentation dynamique du tracé

3.5 Le poser et lever du calame

Dans une même lettre, on peut trouver des points de poser et de lever de calame au même endroit, mais il est difficile de les identifier. Les paléographes utilisent la macrographie pour pouvoir les identifier, par une étude à la loupe et à l'œil nu des déformations des traits. Cette solution consiste à identifier de façon empirique les déformations présentes dans un trait afin de retrouver les points de poser et de lever de plume situés précisément en un même point. Par ce mécanisme, il est possible d'observer les deux actions, lever et poser, en un même point. Afin de simplifier les règles complexes d'exécution des traits, on suggèrera que les posers se trouvent en haut et à gauche et que les levers se situent à droite et en bas, mais cette règle trouve beaucoup d'exceptions, comme c'est le cas sur la figure I.7(b) où pour la formation des traits, on dispose d'un poser et d'un lever de plume qui se situent tous deux à gauche.

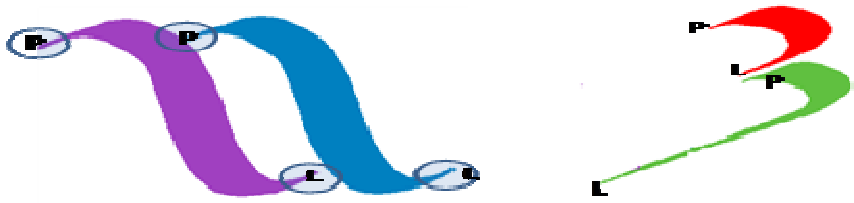


Figure I.7. (a)Poser et lever de calame qui suivent la règle d'exécution («haut-gauche, bas-droite»), (b) cas de poser et lever qui s'écartent de la règle générale

Les ambiguïtés d'interprétation de l'exécution des traits sont, à ce stade, très difficiles à contrôler. Dans ce travail, nous partirons d'hypothèses simplificatrices à partir desquelles nous établirons des règles simples et implémentables régissant la décomposition des traits en graphèmes.

4 Descriptions des bases utilisées

Dans nos travaux nous avons choisi de sélectionner une diversité représentative de manuscrits médiévaux issus de différentes époques, ils sont représentatifs de différents styles d'écriture mais respectent un processus de construction du ductus identique, nous permettant d'appliquer les mêmes hypothèses pour la décomposition des traits en graphèmes. Pour cela nous avons sélectionné des manuscrits de la base d'images paléographiques de l'IRHT de Paris issus de l'époque carolingienne et de l'époque gothique (*310* pages), les manuscrits de l'université d'Oxford dont la diversité graphique est moindre (*140* pages) et enfin, dans un contexte très différent une collection de manuscrits contemporains utilisés dans la compétition d'identification de scripteurs (ICDAR 2011) en anglais, français, allemand et grec. Les manuscrits de cette dernière base ne respectent pas les mêmes règles de composition des traits.

Nous les avons néanmoins soumis aux mêmes études pour montrer la généricité de nos approches qui peuvent s'appliquer sur des manuscrits contemporaints. Les résultats seront présentés au dernier chapitre de la thèse et ouvriront des extensions de nos travaux à des images de contenus plus diversifiés (incluant notamment des images de partitions musicales, de formules, et des images graphiques).

4.1 La base de manuscrits de l'IRHT

La base de l'IRHT comprend *311* manuscrits en niveaux de gris avec une résolution de *300* DPI, partagés en *22* classes allant de l'époque carolingienne à l'époque gothique. La figure I.8 résume le nombre de manuscrits dans chaque classe.

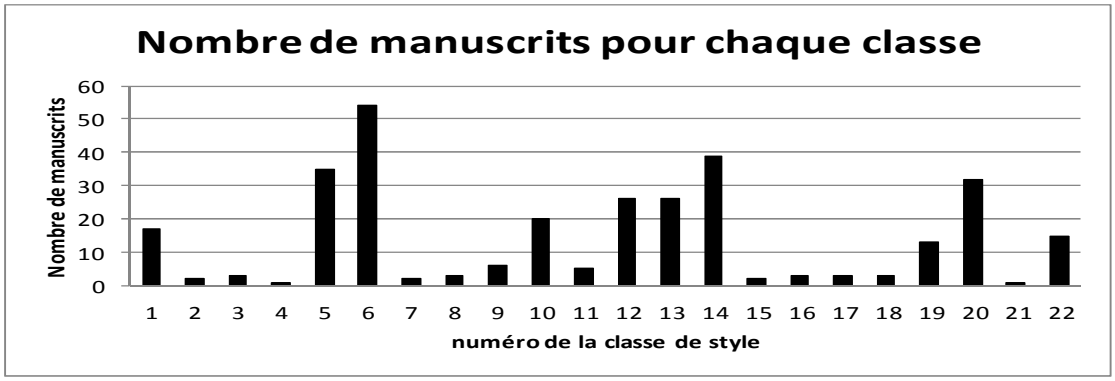


Figure I.8. Classes et nombre de manuscrits constituant la base de manuscrits de l'IRHT

La figure I.9 montre un échantillon des *22* classes de manuscrits. La classification de cette petite base a été produite empiriquement par les experts paléographes afin de guider le processus de description et de classement automatique à mettre en place. Il faut noter que cette tâche n'est pas facile et que les résultats de classification subissent l'influence de plusieurs facteurs (Moalla, [2009]):

- La possibilité d'avoir différentes classifications pour une même écriture, puisque les paléographes peuvent avoir différents points de vue concernant le style d'une écriture. L'hypothèse d'une classification stricte doit donc être écartée.
- L'existence d'une grande variabilité de formes au sein d'un même style d'écriture et d'une ressemblance possible entre les écritures de classes différentes.
- L'absence d'une frontière concrète entre les classes du fait de la présence d'écritures hybrides issues d'un mélange de plusieurs styles.

L'absence de règles paléographiques strictes dans l'établissement des frontières entre classes va nous conduire à proposer une méthodologie paramétrable permettant un partitionnement

flexible en classes d'écriture à partir d'une base de primitives et de descripteurs dont l'importance de pouvoir être pondérée et accessible aux experts eux-mêmes.

Figure I.9. Echantillons de manuscrits représentant les 22 classes de la base réduite de l'IRHT

4.2 Base de manuscrits d'Oxford

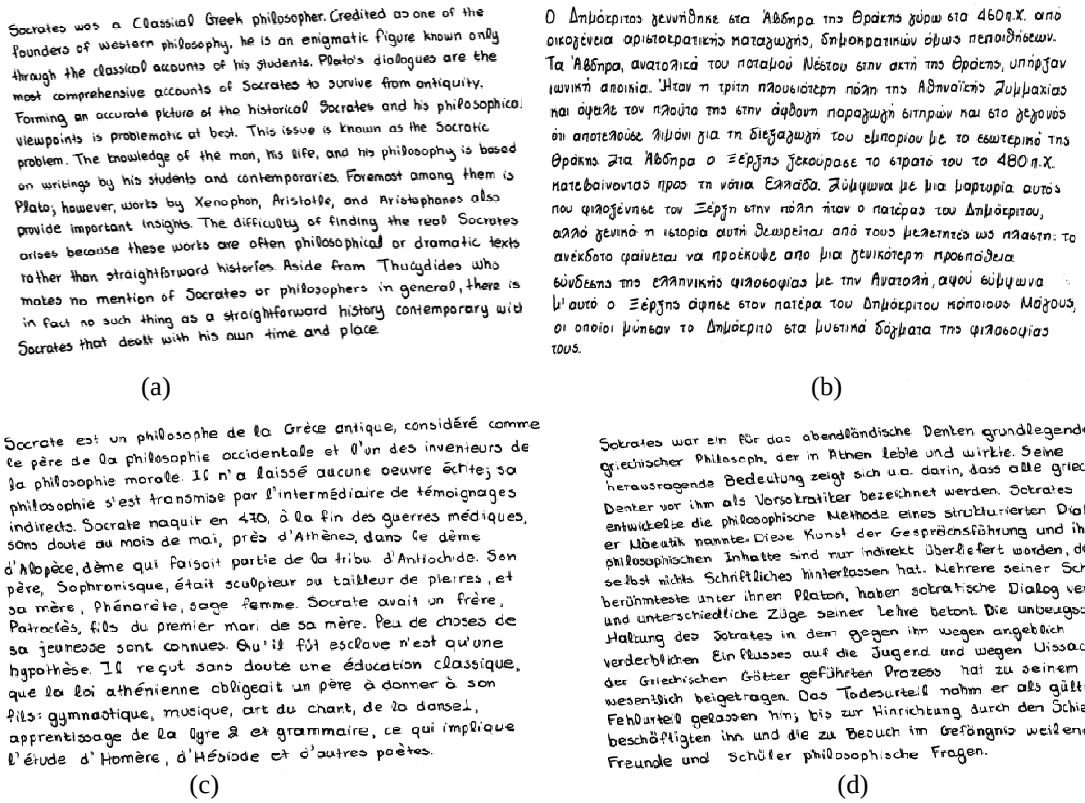
Tout comme la base de l'IRHT, la base de manuscrits d'Oxford[5] est également construite à partir d'images de manuscrits qui résultent de la numérisation de microfilms. La base est constituée de *140* images couleur et qui sont regroupées en *4* classes. Chaque classe est composée de *35* manuscrits. Les manuscrits de la classe *1* sont issus du livre « The Creation of the World » rédigé par William Jordan, *1611*. La classe *2* est composée de manuscrits latins « Sciuias siue Visiones ac reuelatione » appartenant au *12ème* siècle et au début du *13ème* siècle. La classe *3* est une collection de manuscrits latins qui datent de l'année *1372* à l'université de York. La classe *4* est une collection de textes grecs classiques et byzantins. Quelques-uns uniques ou rares, par Michael Psellos et beaucoup d'autres datant du 3ème trimestre du *13ème* siècle. La figure suivante (figure I.10) montre des exemples de chacune des classes.

Figure I.10. Exemples des 4 classes de manuscrits de la base d'Oxford

[5] Université d'Oxford : http://image.ox.ac.uk/list?collection=all

4.3 Base de manuscrits contemporains

Cette base de manuscrits [6] contemporains est constituée de *208* images de manuscrits binarisés avec une résolution de *300* DPI. Elle a été construite avec l'aide de *26* scripteurs à qui on a imposé la consigne de recopier huit pages d'un même texte en plusieurs langues (anglais, français, allemand et grec). Chacune des *26* classes comprend *8* manuscrits. La figure 11 présente quatre manuscrits pour chacune des langues.



(a) (b)

(c) (d)

Figure I.11. Exemple d'un texte produit en quatre langues : (a) anglais, (b) grec, (c) français, (d) allemand, (ICDAR 2011)

L'étude de cette base très particulière sera commentée dans les perspectives de nos travaux : elle ouvre notamment la voie à de nouvelles solutions autour des problèmes de discrimination des graphies selon la langue et le scripteur.

[6] ICDAR 2011 - Writer Identification Contest, http://users.iit.demokritos.gr/~louloud/Writer_Identification_Contest/index.html

Chapitre 1 : État de l’art sur les méthodes de classification et de reconnaissance des styles d’écritures

Résumé : *Dans ce chapitre, nous faisons le point sur les méthodes de classification et de reconnaissance de styles d’écritures qui appartiennent au domaine central de la problématique de ce travail de thèse. Les approches décrites dans ce chapitre portent également sur les aspects d’identification de scripteurs car d’un point de vue méthodologique, certaines techniques peuvent être transverses aux deux domaines (analyse des styles et de scripteurs). Pour chaque approche de classification, nous montrons dans ce chapitre sur quel type de manuscrits l’analyse est appliquée et nous présentons les mécanismes de classification sous-jacents (de la description des contenus: mots, lignes ou pages au processus de classification lui-même, jusqu’à la décision). Nous examinons enfin les méthodes de classification (supervisées, non-supervisées et semi-supervisées) en évaluant pour chacune d’elles les avantages et désavantages et nous montrerons enfin leurs performances dans le domaine de la classification des écritures. La fin du chapitre présente notre choix d’une méthode non-supervisée basée sur la théorie des graphes. Cette approche sera exploitée tout au long de la thèse comme outil de partitionnement privilégié tout à fait adapté au besoin d’un système de classification flexible, simplement paramétrable et exploitable par des non experts du domaine. Le choix de la coloration de graphe est pleinement argumenté dans ce chapitre.*

Mots clés : *Classification supervisée, classification non-supervisée, classification semi-supervisée, coloration de graphe, dictionnaire de formes, style d’écriture.*

1 Introduction

1.1 Les scénarios d’analyse dans le domaine de l’analyse des écritures

L'écriture est un moyen essentiel de communication dans notre civilisation. Elle s'est développée et a évolué au fil du temps. Comme toute production humaine, l'écriture est soumise à de nombreuses variations d'origines très diverses qui pourraient être historique, géographique, ethnique ou sociale. Toutefois, elle a aussi un lien fort avec des caractéristiques innées d'une personne qui expliquent la grande variabilité observée entre les écrits de différents écrivains, même s'ils appartiennent à des communautés voisines. Contrairement à la version électronique ou imprimée, le texte manuscrit comporte des informations supplémentaires sur la personnalité de la personne qui a écrit. Il existe un certain degré de stabilité dans le style d'écriture d'un

individu *(Boulehmi et al. [2008]),* ce qui rend possible le processus d'identification de l'auteur dont on connaît le style.

La classification des écritures, la vérification d'un scripteur ou l'identification d'un auteur constituent des étapes préliminaires aux processus de recherche d'information (recherche de texte selon la similarité d'écritures ou selon l'auteur, partitionnement de grands corpus selon le style…) et d'aide à l'analyse experte (aide à la datation, à l'analyse génétique).

Le classement des écritures et l'identification du scripteur sont des domaines de recherche très actifs durant ces deux dernières décennies. Une grande variété de systèmes est basée sur l'utilisation du traitement d'image par ordinateur et des techniques de reconnaissance de forme ont été proposées pour résoudre les problèmes rencontrés dans l'analyse automatique de l'écriture, voir figure 1.1 inspirée de *(Atanasiu et al. [2011]).* Cette figure décrit le cadre standard de l'analyse des écritures, en distinguant deux points de vue, le point de vue individuel (le scripteur) et le point de vue macroscopique (le style). Différents scénarios d'analyse et de reconnaissance sont ainsi mis en jeu dans ce domaine. Les méthodes d'identification du scripteur ou de classification de style peuvent être regroupées selon la modalité d'acquisition : en ligne et hors ligne. La tâche d'identification en ligne d'un écrivain ou de classification d'un style est généralement considérée comme étant une tâche moins complexe à résoudre que la tâche d'identification hors ligne car elle permet de disposer de davantage d'informations sur le style d'écriture d'une personne, comme la vitesse, l'angle de l'outil d'écriture avec le support ou la pression. Ces données ne sont pas disponibles dans les méthodes hors ligne. Nous n'aborderons dans ce chapitre que les techniques d'identification de scripteurs et de classification des écritures hors ligne.

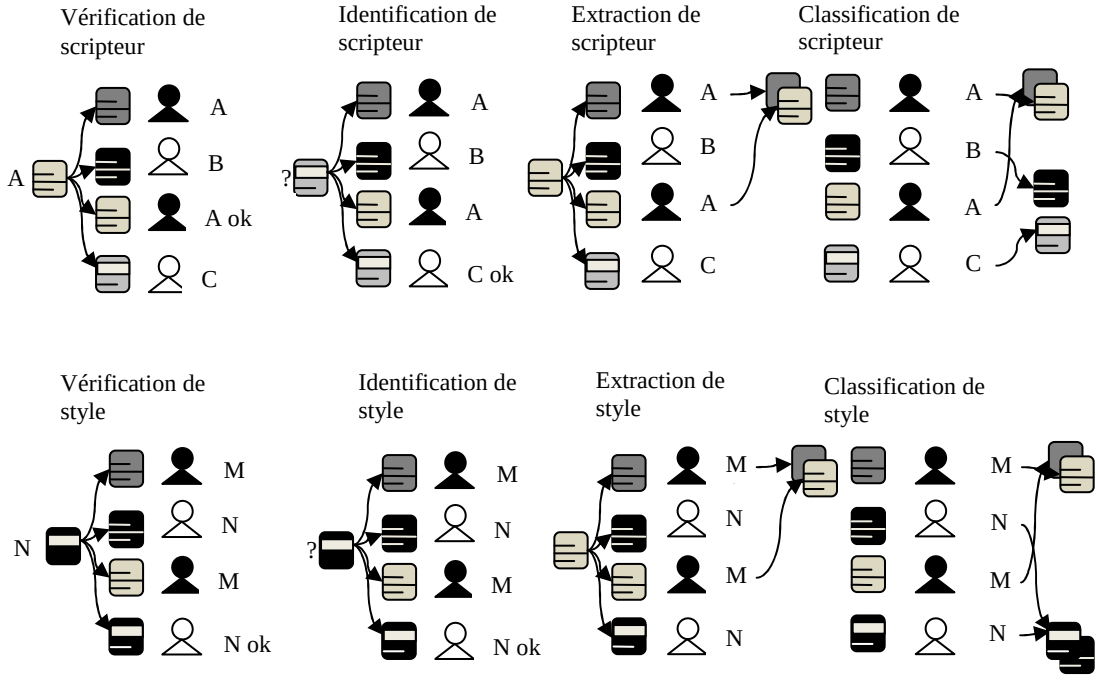


Figure 1.1. Illustration des différents scénarios de classification et de reconnaissance de scripteurs et de styles d'écritures *(Atanasiu et al. [2011])*

1.2 Les étapes fondamentales des systèmes de classification et de reconnaissance de styles

L'identification du scripteur, la classification et la reconnaissance de style d'écriture sont des thématiques de recherche autour des documents qui suscitent beaucoup d'intérêt, en raison de l'existence de grands corpus d'images de documents manuscrits numérisés (patrimoniaux : littéraires, historiques mais également administratifs et techniques). Il est nécessaire d'instrumentaliser ces grands corpus par des techniques d'enrichissement de contenu, la mise en place de méthodes d'accès, de recherche par le contenu et de navigation simplifiée. Cette transformation des corpus est en particulier considérée comme un des défis majeurs dans le contexte des bibliothèques numériques puisque des centaines de milliers de manuscrits de par le monde sont en attente d'être classifiés, indexés et identifiés, *(Bui et al. [2011]), (Bulacu et Schomaker, [2005]), (Bulacu et Schomaker, [2007]), (Brink et al. [2012])*. Le défi est d'autant plus grand que chaque type de ces manuscrits possède des caractéristiques très spécifiques grapho-morphologiques (formation des traits, épaisseur et taille, typicité de la langue et du style) et plus globalement des caractéristiques de mise en forme physique du manuscrit (organisation spatiale et distribution de l'information sur la page). Nous pouvons signaler que

tout système de classification et de reconnaissance de styles repose sur trois étapes fondamentales, celles-ci sont présentées dans la figure 1.2.

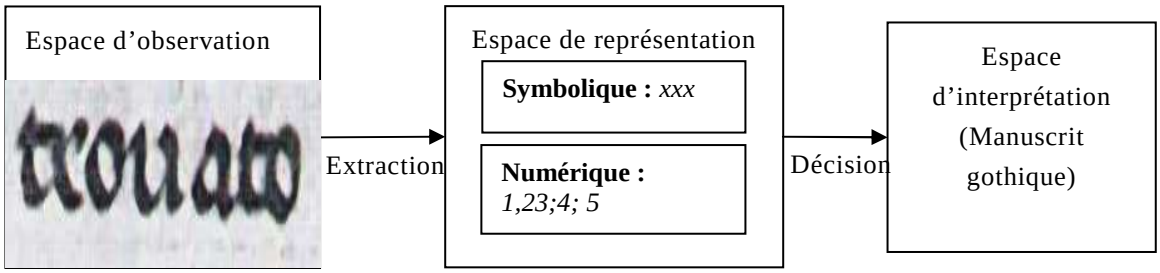


Figure 1.2. Les étapes fondamentales d'un système de reconnaissance de styles d'écriture

En général, l'opération de reconnaissance comprend 2 étapes : l'extraction de caractéristiques (ou de primitives) dont le but est de réduire la quantité de données et d'éviter l'explosion combinatoire et l'opération de classement suivie de la décision qui consiste à associer une description symbolique à l'objet, sur la base de ses caractéristiques. Les caractéristiques sont des valeurs choisies de manière à ce qu'elles soient semblables pour les formes d'une même classe, dissemblables pour des formes de classes différentes.

Il est donc nécessaire de considérer la description des entités manuscrites comme les données d'entrée du classifieur. Cette description peut revêtir différentes formes et porter sur des indicateurs de formes, de textures ou d'organisation structurelle des contenus *(Moalla et al. [2006]), (Joutel, [2009]), (Bulacu et Schomaker, [2005]), (Das et al. [2011]).* L'étape d'extraction de caractéristiques visuelles présentée en détail dans le chapitre 3 consiste en des transformations mathématiques calculées sur les pixels de l'image. Les caractéristiques visuelles permettent généralement de mieux rendre compte de certaines propriétés visuelles des formes en présence.

Dans le cas général, on parle de signatures numériques du manuscrit pour évoquer l'empreinte de son contenu et simplifier sa représentation. Cette signature peut être décrite sous la forme d'un dictionnaire de formes construit à partir des fragments de traits qui constituent l'écriture dans le manuscrit (approche structurelle de *(Bulacu et al. [2005]),* ou à partir de matrices de cooccurrences (approche stochastique de *(Moalla et al. [2006])* ou encore à partir des ondelettes-curvelets (approche fréquentielle de *(Joutel et al. [2008]).* Dans un second temps, après avoir trouvé une représentation pour les manuscrits il est nécessaire de créer des classes de contenus rendant compte des éléments visuellement similaires.

Les étapes de classification des styles et d'identification de scripteurs qui constituent l'objectif de ce chapitre, consistent à associer à chaque manuscrit une classe d'appartenance ou de rejet. Selon le cas, et nous le verrons à travers les éléments bibliographiques évoqués dans

les sections suivantes, le classement peut être formulé comme une liste de classes ordonnées selon une valeur de confiance ou la proposition d'une classe unique solution. Une décision d'identification peut alors être proposée. Lorsque la formulation du problème est telle qu'un résultat binaire est attendu (hypothèse acceptée ou refusée), on parle de vérification.

Parmi les méthodes de classification que nous allons présenter, nous distinguons principalement les méthodes statistiques (l'extraction des caractéristiques produit des valeurs numériques qui sont comparées aux modèles statistiques caractérisant chaque classe), les méthodes structurelles (ou syntaxiques) où l'extraction des primitives produit des valeurs symboliques et des relations qui font l'objet d'une analyse structurelle ou syntaxique, et enfin les méthodes faisant appel à une combinaison de ces deux approches.

L'origine des difficultés associées à tout problème de classification et de reconnaissance est liée en grande partie à la grande diversité des contenus à traiter, la variabilité des formes (figure 1.3) que nous considérons appartenir à une même classe, la présence de distorsions (déformation, bruit) des données. Cela rend la combinatoire dans la résolution des problèmes de reconnaissance plus importante encore.

Figure 1.3. Variabilité du « a » de la base de manuscrits de l'IRHT

La suite de ce chapitre va se dérouler comme suit : dans la prochaine partie nous présentons les méthodes de discrimination et de reconnaissance de styles, discipline centrale à notre problématique. Les techniques seront présentées sous l'angle de la description des contenus balayant un grand nombre de mécanismes portant soit sur une description locale (niveau graphèmes, mots), soit sur une description globale (approche texture). Les techniques de classification que nous avons employées dans notre approche sont également annoncées mais sans détail. Dans la partie suivante du chapitre, nous présenterons les algorithmes de classification qui sont utilisés dans ces méthodes, leurs avantages et leurs limites au regard des problèmes de classification des styles d'écritures. Puis nous argumentons notre approche de

classification non-supervisée basée sur la coloration de graphe, pour enfin terminer par la conclusion.

2 Taxonomie de description de l'information manuscrite pour la discrimination de styles et des scripteurs

La classification des écritures s'appuie sur le fait que chaque manuscrit possède une distribution spatiale unique de ses informations de contenus et des attributs visuels qui lui permettent de le distinguer des autres manuscrits. La tâche de base dans la classification des manuscrits consiste à extraire les caractéristiques d'un manuscrit donné puis à les classifier. Les méthodes proposées seront présentées de la façon suivante : les méthodes basées sur des indicateurs extraits de la page complète, d'un paragraphe court, d'une simple ligne ou d'un mot voire d'une association de fragments (de type graphème). Le choix de la représentation du manuscrit dépend de la taille des échantillons accessibles à partir desquels les caractéristiques extraites pourront être jugées suffisantes et robustes.

Une classification des méthodes selon la modalité d'accès au contenu est présentée dans la section suivante. Elle donne un aperçu du domaine et permet d'apprécier les combinaisons descripteurs/classifieurs selon la nature des contenus et selon deux types d'approches : locales et globales. Les aspects méthodologiques soutenant les approches de classification sont ensuite détaillés (principes, points forts et points faibles des outils de classification et de clustering pris individuellement et ramenés au domaine de la reconnaissance des écritures).

Dans les sections suivantes, nous présentons des travaux portant sur l'analyse et la discrimination de scripts et d'alphabets et sur l'identification de scripteurs. La discrimination d'écritures a connu un franc succès ces dernières décennies en raison notamment de la présence fréquente de plusieurs alphabets au sein d'un même document. Une grande part des travaux cités dans cette revue bibliographique concerne la discrimination des styles d'écritures. Elle illustre la problématique générale de l'analyse et la classification de styles qui constituent le cœur de nos travaux. Dans ce chapitre nous mettons l'accent sur la philosophie générale de chaque méthode sans rentrer dans le détail des descripteurs utilisés. Les différents types de caractérisation feront l'objet d'un état de l'art au chapitre 2.

2.1 Description par connexités : d'une description locale à l'élaboration d'une signature globale de la page

Les manuscrits diffèrent les uns des autres par la structure des traits, les connections qui existent entre ces traits et le style global d'écriture associé à l'ensemble de caractères individuels utilisés. Une approche un peu simpliste mais communément utilisée pour la reconnaissance de l'imprimé consiste à extraire les composantes connexes du texte *(Ghosh et al. [2009], (Ronse et Devijver [1984])* et puis à analyser les formes et leurs structures pointant des caractéristiques essentiellement morphologiques de l'écriture utilisée dans le document. Dans les manuscrits où l'écriture est cursive, les caractères dans un mot ou une partie du mot peuvent se toucher pour former une seule composante connexe. Dans les manuscrits arabes par exemple, un mot ou une partie d'un mot peut former une composante connexe unique. Les méthodes de classification de manuscrits, basées sur l'extraction et l'analyse des composantes connexes appartiennent à la catégorie des méthodes dites locales.

2.1.1 Approches structurelles

Une méthode de classification de manuscrits qui s'appuie sur la relation spatiale des structures internes des caractères a été développée par *(Spitz et al, [1990]), dans ses travaux il a utilisé la densité optique des caractères* pour séparer les manuscrits latins des manuscrits japonais sur les documents imprimés. Dans d'autres travaux portant sur la description structurelle des formes écrites *(Spitz et Ozaki [1994]),* les auteurs ont utilisé la distribution verticale des concavités ascendantes dans les caractères pour produire une discrimination entre le han et le latin avec une précision de *100%.* Il a aussi été proposé un classifieur à deux étapes où sont combinées ces deux caractéristiques : densité optique et distribution verticale. Dans la première étape le han est séparé du latin par comparaison de la variance de la distribution des concavités ascendantes puis dans la deuxième étape du processus de discrimination, la classification des manuscrits han est réalisée à partir de l'analyse de la densité optique du texte. Les travaux présentés par Spitz ont été étendus par *(Lee et al. [1996a])* et *(Waked et al. [1998])* qui ont ajouté plus de caractéristiques pour la description des caractères.

Dans *(Lee et al. [1996a]),* les caractéristiques utilisées sont la distribution des hauteurs du caractère, le profil supérieur et inférieur de la boîte englobant le caractère et exploitant directement les mesures proposées par Spitz (distribution des concavités ascendantes et densité optique). Avec ces caractéristiques, les auteurs ont pu séparer les textes chinois et japonais des textes latins (anglais, français, allemands) dans *98,1%* des cas.

Dans *(Waked et al. [1998]),* les auteurs ont également utilisé la boîte englobant les caractères, la distribution de densité de chaque caractère et la projection horizontale pour

classifier des documents imprimés écrits en japonais, latin, cyrillique et arabe. Ces caractéristiques statistiques sont plus robustes que les caractéristiques basées sur la structure, proposées par Spitz et Lee et al. Mais, en revanche, l'approche de Waked s'est avérée peu robuste aux variations de qualité des documents, aux variations de résolution et de formats. Dans ses travaux il n'est parvenu qu'à une précision voisine de *91%* en testant des documents présentant des particularités variées. Ces performances moindres sont également liées à la ressemblance qui existe entre les manuscrits latins et cyrilliques ainsi qu'aux dégradations altérant les résultats de classification.

2.1.2 Approches statistiques

Dans *(Hochberg et al. [1997]),* les auteurs ont proposé une approche pour la classification de *496* manuscrits arabe, chinois, cyrillique, japonais et latin. L'élément basique de leur analyse est la composante connexe. *5* caractéristiques sont extraites à partir des composantes connexes qui sont les positions relatives verticales des centroïdes, les positions horizontales des centroïdes, le nombre de trous et les ratios entre longueur et largeur des composantes connexes. Pour chaque caractéristique ils calculent la moyenne, l'écart type et le coefficient de dissymétrie formant un vecteur de *15* éléments pour chaque manuscrit. Les auteurs sont parvenus à des résultats atteignant *88%* de bonne discrimination. Un ensemble de classifieurs choisis en fonction du coefficient de discrimination linéaire de Fisher, exploitant un classifieur pour chaque paire de classes a permis d'obtenir ces résultats.

Dans *(Fornes et al. [2008]),* les auteurs ont présenté un système d'identification de scripteur portant sur d'anciens manuscrits de partitions musicales. Le manuscrit est prétraité et normalisé pour obtenir une ligne unique binarisée. Ensuite *100* caractéristiques sont extraites pour chaque ligne, puis sont utilisées dans un classificateur *knn* qui compare chaque vecteur de caractéristiques avec ceux des prototypes stockés dans une base de données. La méthode proposée a été testée sur des manuscrits des $17^{ème}$ et $19^{ème}$ siècles, réalisant un taux de reconnaissance d'environ *95%.*

2.1.3 Approches mixtes

Les dictionnaires de formes

La production de dictionnaires de formes (méthodes également connues sous le terme plus générique de « bag of words ») est apparue comme une méthode très populaire et très efficace pour l'identification d'un scripteur. Le dictionnaire de formes est un tableau présentant le vocabulaire des formes présentes dans les marques d'écriture (et pouvant ainsi être définies soit comme des graphèmes, des caractères, des traits…). Il permet d'associer à chaque entrée le nombre d'occurrences de la forme, dans le document d'étude. Le dictionnaire de formes peut être calculé individuellement pour une écriture ou universellement pour plusieurs scripteurs rassemblant de ce fait l'ensemble des occurrences des formes présentes dans les différents alphabets. Le dictionnaire de formes représente une signature spécifique pour chaque scripteur ou chaque style, et c'est à partir de cette signature qu'il convient de différencier les scripteurs ou les styles entre eux.

Dans *(Schomaker al. [2007]),* les auteurs ont utilisé des dictionnaires de formes conçus à partir des composantes connexes des contours pour l'identification des scripteurs et construits à partir de cartes de Kohonen. La méthode a été testée sur un ensemble de *250* manuscrits à partir de la base Firemaker et à partir de la base Unipen contenant *215* manuscrits avec un taux de reconnaissance de *97%* pour les *TOP-10.*

Dans *(Zhu et al. [2009]),* les auteurs ont utilisé les dictionnaires de formes construits à partir de fragments de contours pour la classification de manuscrits sur une base de *1512* documents contenant respectivement des manuscrits arabes, chinois, anglais, indiens, japonais, coréens, russes et thaïlandais. Le taux de reconnaissance moyen atteint *95,5%.*

Dans *(Ghiasi et Safabakhsh, [2010]),* les auteurs ont utilisé des dictionnaires de formes pour l'identification de scripteurs sur deux bases de manuscrits persans de *40* et *180* manuscrits chacune. Ces deux bases contiennent respectivement des manuscrits de petites, moyennes et grandes tailles. Le taux moyen d'identification atteint *99,5%* pour les manuscrits de grande taille.

Les approches fractales

Une approche de classification de manuscrits basée sur l'analyse fractale a été utilisée pour la discrimination des manuscrits chinois, japonais et devanagari, *(Tho et Tang, [2001]).* Les caractéristiques fractales ont été calculées à partir des signatures des formes extraites des images de manuscrits *(Vincent et al. [2001]),* utilisées aussi pour catégoriser les signatures. La signature fractale est déterminée à partir d'un calcul de surface sur laquelle une fonction de niveau de gris correspondant à l'image du manuscrit est tracée.

Les moments

Une méthode pour la classification des manuscrits arabes et anglais se basant sur l'analyse des composantes connexes a été présentée par *(Elgammal et Ismai, [2001]).* Les caractéristiques extraites sont les projections horizontales et verticales, les sommets de projections horizontales, et les moments des projections horizontales. Les projections de lignes de texte en arabe ont un seul pic vers le milieu de la ligne tandis que les projections de lignes de texte anglais en présentent deux, l'un dans la moitié supérieure de la ligne et l'autre dans la moitié inférieure. Leur méthode a été testée sur les documents imprimés contenant *816* lignes en arabe et *1160* lignes en anglais. Ils ont eu une précision de reconnaissance de *99,7%.*

L'un des premiers systèmes de discrimination de styles d'écriture basés sur l'analyse des caractères a été proposé par *(Lee et Kim, [1995]).* Pour parvenir à cet objectif de catégoriser les manuscrits à partir de quelques caractères, les auteurs ont exploité les réseaux de cartes d'auto organisation. Le réseau mis en place peut identifier le style du manuscrit à partir de chaque caractère du document et le classifier dans un des quatre groupes : latin, chinois, coréen et un groupe mixte selon des descripteurs par moments du premier et du second ordre. Les caractères classifiés dans la classe « mixte » sont en réalité classifiés selon un processus de reclassification plus fin utilisant un apprentissage spécifique par quantification vectorielle. *3367200* caractères ont été utilisés pour mesurer la performance de ce système, et un taux de reconnaissance de *98,27%* a été obtenu.

Dans *(Ablavsky et Stevens, [2003]),* les auteurs ont développé un algorithme d'analyse du flux de composantes connexes et ont attribué une étiquette au texte manuscrit dès que l'accumulation des évidences est jugée suffisante pour prendre cette décision. Cette méthode utilise des propriétés géométriques comme les moments cartésiens et la compacité pour la description des formes. La probabilité que chaque symbole appartienne à une classe de manuscrits donnée, est calculée en utilisant un classifieur *kppv*. Cette approche a produit un taux de reconnaissance de *97%* sur des manuscrits latins et cyrilliques.

Les approches spectrales

D'autres stratégies basées sur des caractéristiques locales extraites à partir des composantes connexes et exploitant la moyenne et l'écart type des sorties des six niveaux du filtre de Gabor ainsi que la distribution du ratio Largeur-Hauteur des composantes connexes ont été développées dans *(Chaudhury et Sheth, [1999]).* Dans les deux cas la classification est appliquée en utilisant la distance de Mahalanobis. Les moyennes de reconnaissance obtenues sur les manuscrits latins, devanagaris, telegu et malayalam sont de *85%, 95%, 98%* et *51%* respectivement.

2.2 Méthodes de discrimination de styles basées sur une description des mots ou des lignes

2.2.1 Approches basées sur les propriétés géométriques des mots

Parallèlement aux travaux portant sur une description du bloc de texte, les premiers travaux concernant la classification de manuscrits portent sur une analyse des lignes et ont été appliqués à l'analyse des manuscrits indiens, urdu, devanagari et dengali. Ces méthodes utilisent la projection du profil, des caractéristiques topologiques et statistiques, et des caractéristiques extraites des traits d'écriture, exploitant ensuite une classification de documents basée sur les arbres de décision.

Dans la première étape du processus de reconnaissance de style présenté par *(Pal et Chaudhuri, [1999]) et (Pal et Chaudhuri, [2002]),* les descripteurs de formes sont extraits sur les zones de titre pour séparer les pages bengari et devanagari du latin, chinois et arabe. Puis la séparation des entités manuscrites bengari et devanagari est réalisée en observant la présence de traits spécifiques à chaque style. La séparation entre les lignes de texte chinois est effectuée en vérifiant l'existence de caractères avec au minimum quatre traits verticaux.

Finalement la séparation des lignes de texte latin (anglais) des lignes de texte arabe est achevée en utilisant des caractéristiques statistiques qui incluent la distribution des points les plus bas dans les caractères. Les points les plus bas dans les lignes de textes latins se trouvent tout le long de la ligne de base la plus haute tandis que les points dans les lignes de textes arabes sont distribués aléatoirement, de même les caractéristiques structurales basées sur la concavité des caractères ont été utilisées. Le taux moyen obtenu est de *97,33%* pour des manuscrits latin, chinois, arabe, bengali, devanagari.

Comparée aux processus de discrimination de styles d'écritures et de scripts basés sur les blocs de texte et les lignes, la classification de scripts portant sur l'analyse des mots (et plus encore des caractères) est généralement plus difficile. Ceci est lié au fait que la quantité d'information disponible à partir de quelques caractères présents dans un mot peut ne pas être suffisante pour l'application.

Deux autres méthodes de séparation des mots arabes et anglais ont aussi été proposées. Dans la première *(Moalla et al. [2002]),* une base d'apprentissage contenant des modèles de segments de caractères arabes est générée. Un mot est supposé être en arabe si le pourcentage de correspondance du segment dans le mot dépasse une valeur définie par l'utilisateur, sinon, le mot est considéré comme étant écrit en anglais. Les expériences ont abouti à un taux de reconnaissance de *100%* sur *30* textes contenant *478* mots.

Dans la seconde méthode basée sur la reconnaissance des mots arabes, la correspondance de caractéristiques a été utilisée *(Moalla et al. [2004]).* Les caractéristiques utilisées sont des caractéristiques morphologiques et statistiques telles que le chevauchement et l'inclusion des boîtes englobantes, une barre horizontale, les signes diacritiques bas, hauteur et largeur de la variation des composantes connexes. La précision de la reconnaissance obtenue avec cette méthode était de *98%.*

2.2.2 Approches basées sur les propriétés spectrales des mots

Dans les travaux de *(Ma et al. [2003]) (Jaeger et al. [2005])* le filtre de Gabor est appliqué sur chaque mot dans un document bilingue pour extraire des caractéristiques caractérisant le manuscrit dans lequel ce mot particulier est écrit. Un classifieur à deux classes a été utilisé pour discriminer ces deux types de langues présentes dans le manuscrit. Des architectures différentes de classifieurs ont été considérées : *SVM, kppv,* distance euclidienne pondérée ainsi qu'un modèle de mélange gaussien. Un système avec un seul classifieur peut contenir un des quatre classifieurs cités avant et un système de classifieurs multiples est construit en combinant deux ou plusieurs classifieurs. Dans le système de classifieurs multiples, les résultats de classification de chacun des classifieurs inclus dans le système sont combinés en utilisant une sommation des scores pour arriver à la décision finale. Dans *(Ma et al. [2003]),* les auteurs ont comparé en particulier des manuscrits anglais et des manuscrits arabes, chinois, indiens et coréens. La performance de classification des manuscrits anglais et indiens utilisant les *kppv* est de *97,51%* tandis que la classification des manuscrits anglais et arabes avec les classifieurs *SVM* est réalisée avec un taux de *91%.* Dans leurs travaux, les auteurs ont pu prouver qu'un ensemble de classifieurs peut considérablement améliorer la performance de classification atteignant des taux de *98%* pour la classification des manuscrits anglais, indiens et *92,66%* pour celle des manuscrits anglais, arabes, et cela par combinaison des classifieurs *kppv* et *SVM.*

D'autres approches, portant sur les propriétés spectrales des mots dans un environnement bilingue pour la classification des manuscrits tamil et anglais ont également été proposées. Nous allons citer les deux approches suivantes *(Dhanya et al. [2002]):* La première méthode structure les mots en trois zones distinctes spatiales et utilise l'information de la répartition spatiale des mots dans ces zones. La seconde approche analyse la distribution de l'énergie directionnelle des mots en utilisant les filtres de Gabor avec des fréquences et orientations adéquates.

Les algorithmes sont basés sur les observations suivantes : la répartition spatiale des caractères latins sont répartis dans les zones moyennes et supérieures, seuls quelques caractères minuscules sont répartis dans la zone inférieure, l'alphabet latin contient plus de traits verticaux et horizontaux. Dans le tamil, les caractères sont répartis dans les zones supérieures et inférieures, il existe une dominance des traits horizontaux et verticaux, le rapport entre la

hauteur et la largeur des caractères tamils est généralement plus grand que celui des caractères latins. Ces résultats suggèrent que les caractéristiques qui peuvent jouer un rôle majeur dans la discrimination des manuscrits latins et tamils sont la répartition spatiale des mots et le sens de l'orientation des éléments structurels des caractères dans les mots. Les caractéristiques spatiales sont obtenues en calculant la concentration des pixels dans chacune des trois zones, tandis que les caractéristiques directionnelles sont calculées avec les filtres de Gabor.

Les caractéristiques extraites sont classifiées en utilisant les *SVM* ou les *kppv*. Il a été observé que les caractéristiques directionnelles possèdent de meilleures capacités de discrimination que les caractéristiques spatiales, et ont produit une précision de *96%* en utilisant un classifieur *SVM*. Cela peut être attribué au fait que les filtres de Gabor prennent en compte la nature générale des manuscrits mieux que la répartition spatiale.

2.3 Méthodes de discrimination de styles basées sur une description texture des textes

Alors que les approches précédentes étaient basées sur une analyse de connexités (ramenée à une description finalement macroscopique d'un manuscrit par cumul des descriptions individuelles), les méthodes de classification de manuscrit discutées dans cette section nécessitent des informations plus étendues de la taille du bloc ou du paragraphe de texte. Les approches précédentes nécessitent la présence d'un minimum d'information pour être performantes. La présence de quelques connexités réparties sur de petits blocs de texte ne suffit pas toujours à produire des résultats de qualité.

Dans les approches que nous présentons ici, la description repose sur une analyse plus globale portant directement sur des régions de texte de taille significative (bloc et paragraphe).

Certaines de ces approches peuvent également être identifiées sous le label « approche par analyse de texture ».

Une première tentative pour caractériser un manuscrit sans réellement analyser la structure des composantes connexes qui le constituent a été faite par *(Wood et al. [1995]).* Ils ont proposé d'utiliser les projections verticale et horizontale de l'image de manuscrit afin de l'identifier. Ils considèrent que les profils des histogrammes de projections suffisent pour classifier le manuscrit. Par exemple les manuscrits latins montrent des pics en haut et en bas du profil de projection horizontale tandis que dans les manuscrits cyrilliques la ligne médiane est dominante. Les manuscrits coréens présentent des pics à gauche du profil vertical. Toutefois, les auteurs n'ont pas suggéré la manière d'analyser automatiquement ces profils de projection pour le manuscrit sans intervention de l'utilisateur. De plus, aucun résultat de reconnaissance n'a été proposé pour appuyer leurs arguments.

Comme l'apparence visuelle est souvent liée à la texture, un bloc de texte appartenant à chaque classe de manuscrit forme un modèle de texture différent. Le problème d'identification du manuscrit se ramène alors à un problème d'analyse de texture où toute méthode d'analyse de texture peut être utilisée pour effectuer cette tâche.

Dans *(Tan et al. [1998]),* les auteurs ont développé une solution portant sur une fonction de Gabor pour l'analyse de la texture des documents imprimés. La méthode a donné une précision de *96,7%* en discriminant les documents imprimés chinois, latin, grec, russe, persan et malayalam. Dans la première partie de la méthode, un bloc de texte uniforme sur lequel l'analyse de textures est appliquée a été produit. Les caractéristiques ont ensuite été extraites des blocs de texte en exploitant un filtrage de Gabor avec une fréquence radiale fixe de *16* cycles/seconde et *16* orientations équidistantes. Durant la classification un vecteur de caractéristiques est comparé aux vecteurs représentatifs de chaque classe en utilisant une distance euclidienne pondérée. Un vecteur de caractéristiques représentatif pour la classe d'un manuscrit est obtenu en calculant le vecteur de caractéristiques moyen obtenu à partir d'une grande base d'apprentissage de manuscrits.

Un inconvénient de cette méthode, est que les blocs de texte extraits des documents n'ont pas nécessairement un espacement uniforme des caractères. Pour contourner ce problème dans *(Peake et Tan, [1998]),* les auteurs ont amélioré la méthode en utilisant des algorithmes simples de prétraitement d'image permettant d'obtenir des blocs de texte uniformes, comme par exemple la suppression des lignes de texte hors norme et la normalisation de l'espacement inter ligne. Les matrices de cooccurrence et des filtres de Gabor à multiples canaux ont été utilisés indépendamment dans les expériences. Ces deux approches utilisées pour l'extraction des caractéristiques à partir des textures ont été appliquées sur des documents imprimés. La classification des manuscrits a été ensuite réalisée en utilisant un classifieur *kppv*. L'approche portant sur l'exploitation des matrices de cooccurrence a conduit à une précision de 74% alors que celle basée sur les filtres de Gabor a produit une précision de *95,8%.*

Un des problèmes lié à l'usage des filtres de Gabor est le coût élevé de calcul dû aux fréquentes opérations de filtrage de l'image. Pour résoudre ce problème, l'utilisation des filtres de Gabor orientables a été proposée par *(Pan et al. [2005]).* Cette méthode offre deux avantages : Tout d'abord, la propriété d'orientation inhérente aux filtres de Gabor réduit le coût de calcul. Les filtres de Gabor ont été conçu de telle sorte que les caractéristiques extraites du texte et invariantes à la rotation puissent discriminer les manuscrits contenant des caractères similaires en forme et partageant des propriétés morphologiques. Dans *(Pan et al. [2005]),* les auteurs présentent des résultats atteignant une précision de *98,5%* pour la discrimination des manuscrits chinois, japonais, coréens et latins.

L'utilisation des caractéristiques de texture pour la classification des manuscrits a été considérée par *(Jain et Zhong, [1996])* pour discriminer les documents imprimés chinois et anglais. Dans leurs travaux, les auteurs ont proposé un algorithme de segmentation basé sur la texture. Le texte est extrait ainsi que les régions d'images et les zones de dessins à partir des images en niveaux de gris.

L'utilisation des caractéristiques basées sur la texture autres que les matrices de cooccurrence et le filtre de Gabor a été proposée par *(Busch et al. [2005]).* Les caractéristiques utilisées sont les caractéristiques énergétiques d'ondelette, les caractéristiques moyennes de déviations d'ondelette, et des caractéristiques logarithmiques de cooccurrences de coefficients d'ondelettes. Ces caractéristiques ont été utilisées sur une base de test contenant huit types de manuscrits : latin, han, japonais, grecque, cyrillique, hébreux, devanagari et persan. Dans les expériences réalisées, des opérations de prétraitement ont été appliquées sur les documents imprimés produisant des blocs de texte normalisés à des carrés de *64x64*. Pour réduire la dimension des vecteurs de caractéristiques et augmenter la performance du classifieur, le classifieur discriminant linéaire de Fisher a été utilisé. La classification est appliquée en utilisant un modèle de mélange de gaussiennes qui représente chaque manuscrit comme une combinaison de distributions gaussiennes. Les caractéristiques logarithmiques des valeurs de cooccurrences d'ondelettes donnent les meilleurs résultats de classification. Les matrices de cooccurrences ont fourni les taux les plus bas de reconnaissance avec *9,1%* de taux d'erreur. Cela indique que les relations entre les pixels sur de petits voisinages sont insuffisantes pour caractériser un manuscrit correctement.

Dans les travaux de *(Moalla et al. [2009])* les matrices de cooccurrences paramétriques multi-échelles ont été utilisées pour classifier des manuscrits médiévaux. Dans ces travaux, la classification est appliquée sur deux niveaux : la première classification est faite sur les manuscrits carolingiens, pré-gothiques, gothiques et humanistiques sans tenir compte des sous-familles de styles avec un taux de reconnaissance moyen de *97,8%.* Au deuxième niveau, les sous-familles de styles sont prises en compte et la classification est appliquée sur les manuscrits carolingiens, pré-gothiques, gothiques comprenant les sous-familles, textualis, hybrida, cursiva, rotunda et humanistiques avec un taux de reconnaissance moyen de *96%.*

Dans *(Joutel et al. [2008]),* les auteurs ont utilisé la décomposition par curvelets des images de manuscrits. Les curvelets sont des transformées fréquentielles des images qui appartiennent à la famille des ondelettes et qui ont été utilisées à différents niveaux d'échelles sur des manuscrits médiévaux et sur des manuscrits du $18^{ème}$ siècle. La signature produite à partir de l'extraction de la fréquence d'apparition de pixels de certaines orientations et courbures a pu conduire à une description globale des images de textes. En rendant compte des orientations privilégiées et de leurs courbures associées, les signatures des textes ont permis d'obtenir un

taux de bonne classification de *96,61%* sur une base réduite de *300* images de documents médiévaux de l'IRHT et sur une base constituée de manuscrits latins de l'époque des Lumières.

2.4 Bilan sur les méthodes d'identification de manuscrits

Nous présentons dans cette section un bilan des méthodes de discrimination de styles basées sur l'analyse de la connexité, l'analyse des mots, des lignes et l'analyse de texture. Dans un premier temps, nous constatons que le choix des caractéristiques extraites des manuscrits affecte les performances du classifieur: dans *(Tan, [1998])* par exemple, l'utilisation du classifieur *kppv* portant sur des caractéristiques basées sur le filtre de Gabor a conduit à un taux de reconnaissance de *96 %*. Ces valeurs sont à comparer aux résultats obtenus à partir de caractéristiques basées sur les matrices de cooccurrence de niveaux de gris qui n'ont permis d'atteindre que des taux de reconnaissance très réduits de *7,2%*.

Pour les méthodes basées sur l'analyse du mot, nous constatons que les performances diffèrent d'une langue à une autre malgré un usage des mêmes caractéristiques. Dans *(Moalla al. [2002])* des caractéristiques morphologiques et statistiques ont été extraites à partir de deux bases (latine, anglaise et arabe) et ont fourni des taux de reconnaissance de respectivement *100%* et *98%*. La plupart des méthodes d'analyse de manuscrits sont sensibles au type de manuscrit traité. Dans la méthode de *(Roy et Vajda, [2005])* des caractéristiques morphologiques sont extraites à partir des chiffres Bengali imprimés et manuscrits : les taux de reconnaissance sont respectivement de *98,5%* et *89%* selon la nature des contenus (imprimés et manuscrits).

A partir des méthodes d'identification de scripteurs et de classification de style d'écriture que nous avons illustrées dans la section précédente et qui utilisent différentes combinaisons de caractéristiques et d'algorithmes de classification, nous constatons que le choix des caractéristiques et des algorithmes de classification est très important. Pour cela dans la section suivante nous étudions les aspects fondamentaux des algorithmes de classification afin de choisir l'algorithme qui convient le plus à notre problématique. Le choix argumenté des descripteurs que nous avons retenus sera présenté au chapitre *2*.

3 Les algorithmes de classification supervisés et non-supervisés (Catégorisation de manuscrits en styles)

Traditionnellement nous distinguons trois approches de classification: les approches de classification supervisées, les approches de classification non-supervisées et les approches de classification semi-supervisées (Figure 1.4).

- **Classification supervisée :** En mode supervisé, l'objectif est d'apprendre, en fonction de leur description, à classer des exemples dans l'une des n classes, n étant fixé a priori. Pour cela, à partir d'un échantillon d'objets correctement étiquetés, on désire construire une fonction capable d'étiqueter et de classer au mieux de nouveaux objets ne faisant pas partie de l'échantillon initial. Un modèle de classification est donc appris à partir d'un ensemble d'exemples étiquetés (phase d'apprentissage), le but étant de permettre à l'apprentissage d'induire un fort pouvoir de généralisation aux étapes de classification.
- **Classification non-supervisée :** En mode non supervisé, les exemples fournis ne sont pas étiquetés. L'objectif est de les regrouper en catégories cohérentes selon certains critères.
- **Classification semi-supervisée :** La classification semi-supervisée constitue un cas intermédiaire entre l'apprentissage supervisé et non supervisé. Cette technique de classification utilise un ensemble de données étiquetées et non-étiquetées mais le nombre de classes est connu a priori pour faire la classification.

Il existe aussi d'autres approches de classification comme par exemple la classification transductive introduite par *(Vapnik, [1998]).* Cette approche considère généralement une structure de graphe dont les sommets représentent les données et dont le but est de propager l'information des exemples étiquetés sur l'ensemble du graphe. Une fois cette fonction apprise, les exemples non étiquetés de la base sont ordonnés en utilisant les valeurs de scores ainsi obtenues *(Truong et Amini, [2008]).*

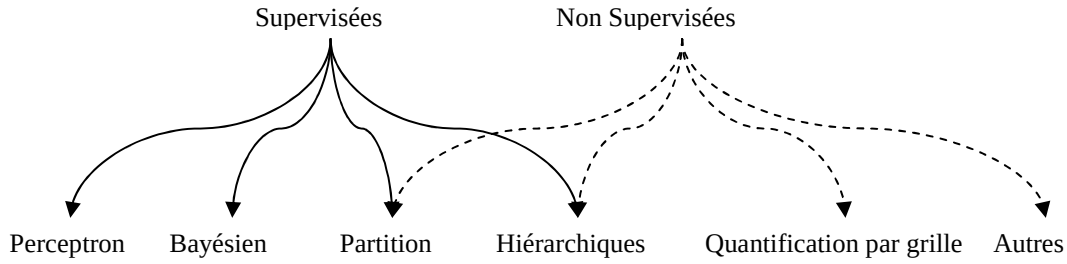


Figure 1.4. Représentation des deux familles de classification supervisée et non-supervisée

Les méthodes suivantes seront illustrées par les résultats exprimés en taux de bonne classification, se référant aux différentes contributions présentées en section 2 de ce chapitre.

3.1 Les méthodes supervisées

Dans cette section nous illustrons les méthodes de classification supervisées. Pour chaque méthode nous allons fournir le principe, les avantages et inconvénients et montrer à partir d'un tableau, l'application de ces méthodes dans le domaine de l'analyse des styles ou l'identification des scripteurs selon le cas. Dans chaque tableau nous montrons les caractéristiques utilisées, le type de document sur lequel les travaux ont porté et le taux de reconnaissance qui a pu être atteint.

3.1.1 Les méthodes de type partitionnement

Dans cette partie, nous présentons les méthodes de classification supervisée de type partitionnement nécessitant un apprentissage des paramètres du modèle permettant de conduire la classification. Des plus simples (l'approche intuitive par k plus proches voisins (*kppv*)) aux plus complexes comme les méthodes à noyaux, nous montrerons les atouts de ces méthodes dans un cadre limité à la classification des écritures, des styles et d'identification de scripteur (selon le cas).

Les *k* Plus Proches Voisins

L'algorithme des *kppv* est basé sur le principe que chaque instance de la base peut être associée à d'autres instances en fonction de propriétés similaires *(Cover et Hart, [1967]).* Si les instances sont étiquetées, alors la valeur de l'étiquette d'une nouvelle instance à classifier est déterminée par un calcul de similarité avec les autres instances. L'algorithme des *kppv* localise les *k* plus proches instances de l'instance considérée comme requête et identifie sa classe en fonction de celles des *k* plus proches instances étiquetées. En général, les instances sont considérées comme des points dans un espace à n-dimensions où chacune des n-dimensions correspond à une des n-caractéristiques qui sont utilisées pour décrire une instance.

Généralement, la distance entre instances est déterminée en utilisant l'une des distances suivantes : Manhattan, Minkowsky, Euclidienne, Camberra, Chebychev. La métrique choisie doit minimiser la distance entre deux instances de la même classe, alors qu'elle maximise la distance entre deux instances de classes différentes.

La puissance des classifieurs de type *kppv* a été démontrée dans plusieurs cas réels *(Phyu, [2009])* mais il reste toujours des réserves liées à la sensibilité de la méthode au choix de la fonction de similarité utilisée pour comparer les instances, à la difficulté de choisir la bonne valeur de k. Sauf à travers la validation croisée ou des techniques de calcul coûteuses, il faut noter que les choix de k affectent énormément la performance de l'algorithme *kppv*. Comme le processus est transparent, les *kppv* sont faciles à mettre en œuvre, dans les situations où une explication de la sortie du classificateur est utile, grâce à une analyse des voisins. Il existe des techniques de réduction du bruit qui ne fonctionnent que pour les *kppv* et qui peuvent être efficaces dans l'amélioration de la précision du classificateur. En revanche, comme tout l'ensemble des calculs se réalisent de l'exécution, les *kppv* peuvent avoir de mauvaises performances réduites à l'exécution si l'ensemble d'apprentissage est grand, de même ils sont sensibles aux caractéristiques non pertinentes et redondantes puisque toutes les caractéristiques contribuent à la classification. Cela peut être amélioré par une étape de sélection ou de pondération des caractéristiques *(Cunningham et Delany, [2007])*.

Le tableau 1.1 illustre le bilan sur les méthodes d'identification de scripteurs basées sur les *kppv*. Nous remarquons que le type de caractéristiques utilisées affecte le taux de reconnaissance du classifieur, *(Peake et tan, [1998])* ont utilisé deux types de caractéristiques sur les mêmes manuscrits : « Les matrices de cooccurrence et les filtres de Gabor » avec une différence de *18,57%*. Le choix des caractéristiques est donc très important et la bonne adéquation entre le choix des caractéristiques les plus adaptées et le type de documents est un élément essentiel influençant très fortement les taux de reconnaissance.

Tableau 1.1. Résultats du classifieur par *kppv* pour la reconnaissance de styles d'écriture

Auteur(s)	Caractéristiques	Type de document	Taux de reconnaissance
(Ablavasky et Stevens, [2003])	Moments cartésiens, compacité	latin, cyrillique	*97%*
(Peake et tan, [1998])	Matrice de cooccurrences	chinois, grecque, latin, malayalam, russe, persan, coréen	*77,14%*
	Filtre de Gabor		*95,71%*
(Joshi et al. [2006])	Filtre de Gabor, projection horizontale du profil, ratio des énergies	devnag, latin, gurumukhi, kannada, malayalam, urdu, tamil, gujrati, oriya, beng.	*97,11%*
(Rajput et Anita, [2010])	Transformée en Cosinus Discrète, ondelette	anglais, kannada, tamil, telagu, gujarati, indien	*90% à 99,2%*
(Das et al. [2011])	Ligne supérieure max, ligne inferieure max, lignes horizontales, lignes verticales…	telagu, anglais, indien	*91,05% à 99,01%*
(Paretiet Vincent. [2008])	Modèle de Loi de Zipf	latin	*62% à 80%*

Machines à vecteurs de support (*SVM*)

Les *SVM* sont considérées comme les plus récentes techniques utilisées dans le domaine de l'apprentissage automatique, elles ont été introduites par Vladimir Vapnik, *(Vapnik, [1995]).* Parmi les méthodes à noyaux, les *SVM* constituent la forme la plus connue. Il s'agit d'une méthode de classification binaire par apprentissage supervisé. Les *SVM* évoluent autour de la notion de « marge » de chaque côté d'un hyperplan qui sépare deux classes de données. Maximiser la marge, créant ainsi la plus grande distance possible entre l'hyperplan séparateur et les instances de chaque côté de celui-ci, garantit la meilleure réduction de l'erreur de généralisation attendue. Puisqu'il s'agit d'un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage supervisées pour apprendre les paramètres du modèle.

Si les données d'apprentissage sont linéairement séparables, alors une paire *(w,b)* existe telle que :

$$\begin{aligned} w^T x_i + b \geq 1 &\text{ , pour tout } x_i \text{ de la première classe} \\ w^T x_i + b \leq -1 &\text{ , pour tout } x_i \text{ de la seconde classe} \end{aligned} \qquad \textbf{(1.1)}$$

Avec la règle de décision donnée par $f_{w,b}(x) = \text{sgn}\ (w^T x + b)$ où *w* est appelé le vecteur de poids et b le biais (ou - b est appelé le seuil).

Il est facile de montrer que, lorsqu'il est possible de séparer linéairement deux classes, un hyperplan séparateur optimal peut être trouvé en minimisant le carré de la distance des éléments étiquetés à l'hyperplan séparateur. La maximisation peut être configurée comme un problème de programmation quadratique:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ contidionné par } \; y_i(x_i \cdot w + b) - 1 \geq 0. \qquad \textbf{(1.2)}$$

Dans le cas de données linéairement séparables, une fois l'hyperplan séparateur optimal trouvé, les points se trouvant sur sa marge sont appelés vecteurs de support et la solution est représentée comme une combinaison linéaire de ces seuls points (figure 1.5). Les autres points des données sont ignorés.

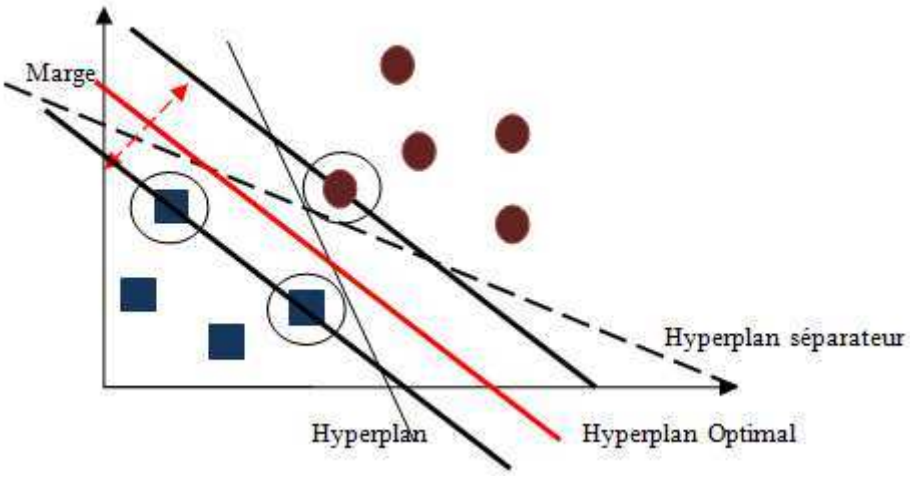


Figure 1.5. Hyperplan à marge-maximale et des marges pour un *SVM* formé avec des échantillons de deux classes. Les échantillons sur la marge sont appelés les vecteurs supports

La complexité du modèle *SVM* n'est pas affectée par le nombre de caractéristiques rencontrées dans les données d'apprentissage. Pour cette raison, les classifieurs *SVM* sont bien adaptés pour faire face aux tâches d'apprentissage où le nombre de fonctions est important par rapport au nombre d'instances d'apprentissage, mais le problème des *SVM* c'est qu'ils conduisent à des méthodes binaires. Ainsi, dans le cas d'un problème multi-classes, il faut réduire le problème à un ensemble de problèmes de classification binaire. Une classification par paire peut être utilisée (une classe contre toutes les autres, et cela pour toutes les classes). Les *SVM* produisent des classifieurs très précis et peuvent traiter un grand nombre de paramètres de manière efficace tout en limitant le sur-apprentissage *(Vapnik, [1995]).* Ils sont robustes au bruit et ont une rapidité d'apprentissage en relation avec le nombre d'attributs et le nombre d'instances.

Dans le tableau 1.2 nous illustrons le bilan sur les méthodes d'identification de scripteurs basées sur des *SVM*. Ici encore, le type de caractéristiques utilisées va avoir un impact sur les résultats. Dans *(Das et al. [2011])* et *(Chanda et al. [2004])* les auteurs ont travaillé sur le même type de documents (bengla) mais en utilisant des caractéristiques différentes ce qui a affecté les taux de reconnaissance, la différence entre les deux approches est de *18%*. Cela illustre une réalité que nous défendons : l'importance du choix des caractéristiques dans les performances d'une classification.

Tableau 1.2. Résultats des *SVM* sur la reconnaissance de styles d'écritures

Auteur(s)	Caractéristiques	Types de document	Taux de reconnaissance
(Ma et Doermann, [2003])	Filtre de Gabor	latin, indien	*90,93% à 92,66%*
(Dhanya et al. [2002])	Caractéristiques directionnelles basées sur le filtre de Gabor	latin, tamil	*96%*
(Kumar et al. [2003])	Caractéristiques à partir du squelette:intersection, point de fin de trait	gurmukh	*94,29%*
(Camastra et Vinciarelli, [2001])	Ratio Hauteur/Largeur, Moments de Zernike	anglais	*90,05%*
(Das et al. [2011])	Projection	bangla	*80,51%*
(Chanda et al. [2004])	Chaîne de Freeman, gradient	anglais, devanagari, bengla	*98,51%*
(Imdad et al. [2007])	Filtre Hermite (Martens, [1990])	anglais	*83%*

3.1.2 Les méthodes hiérarchiques

Les arbres de décision

Les arbres de décision prennent comme entrée un objet ou une situation décrite par un ensemble d'attribus continus ou discrèts et retourne une décision. La valeur de sortie peut être aussi continue ou discrète. L'apprentissage d'une fonction discrète permet une classification et l'apprentissage d'une fonction continue permet une régression.

Les arbres de décision fonctionnent d'une manière récursive. Tout d'abord, un attribut doit être sélectionné comme nœud racine. Afin de créer l'arbre le plus efficace, le nœud racine doit diviser efficacement les données. Chaque étape divise un ensemble d'instances (les données réelles) jusqu'à ce que les instances d'un ensemble aient toutes le même classement. La meilleure division est celle qui fournit ce qu'on appelle le plus grand gain d'informations. Ce gain est calculé à partir de l'entropie.

Pour les méthodes de classification citons l'algorithme C4.5 *(Quinlan, [1993])* qui est une extension de l'algorithme ID3 *(Quinlan, [1979])*. Parmi les méthodes basées sur la régression nous citons l'algorithme CART. Les arbres de régression sont similaires aux arbres de

classification, à l'exception que chaque feuille de l'arbre produit un nombre réel *(Kohavi et Quinlan, [2002]).*

Les arbres de décision sont compréhensibles par l'utilisateur, cela les rend accessibles et appréciés (si la taille de l'arbre produit est raisonnable). Ils permettent d'avoir une traduction immédiate en termes de règles de décision.

Les méthodes liées à la construction des arbres de décision sont non optimales : les arbres produits ne sont pas les meilleurs. En effet, les choix dans la construction des arbres, basés sur de nombreuses heuristiques, ne sont jamais remis en question (pas de retour en arrière ou backtraking) ce qui les rend très sensibles au bruit présent dans les données, des arbres flous ont été introduits pour amoindrir ce défaut. Enfin, il est possible de modifier les valeurs de nombreux paramètres, de choisir entre de nombreuses variantes. Ce lourd paramétrage peut être considéré comme un frein à leur usage. Enfin, la taille de l'ensemble des échantillons influera sur les critères d'élagage à choisir (ensemble d'apprentissage, ensemble de test, validation croisée, ...).

Le tableau 1.3 illustre les résultats obtenus par les méthodes basées sur des arbres de décision dans la classification de manuscrits. Nous pouvons dire que ces méthodes offrent de bons taux de reconnaissance, mais encore, comme nous l'avons remarqué précédement, le choix des caractéristiques joue un rôle important dans l'identification du scripteur. Dans ce tableau, « le cross count » représente le nombre moyen d'intervalles noirs qui se trouvent dans les huit lignes de balayage consécutives qui traversent une image bitmap soit horizontalement, soit verticalement.

Tableau 1.3. Résultats des Arbres de décision sur la classification des manuscrits

Auteur(s)	Caractéristiques	Type de document	Taux de reconnaissance
(Lin et al. [2011])	Densité,concavité, ratio, cross count	chinois, anglais	*88,65% à 99,45%*

3.1.3 Les approches bayésiennes

Les méthodes de décision bayésienne

La décision bayésienne est une méthode qui attribue des fonctions discriminantes g_i $(i = 1,2,\ldots, r)$ respectivement aux classes w_1 $(i = 1,\ldots, r)$. Pour un motif x, nous décidons que x appartient $à$ w_1 si $gi(x) > gj(x)$ (pour tout $j \neq i$) *(Funahashi, [1998]).* Comme fonctions discriminantes g_i $(i = 1,2,\ldots, r)$, on utilise $g_i(x) = p(w_1|x)$: une densité de probabilité a posteriori pour x.

Il est possible d'utiliser $f(p(w_i|x)) + h(x)$ comme $g_i(x)$ où f est une fonction monotone croissante et $h(x)$ une fonction linéraire, ce qui laisse les résultats de classification inchangés. Par exemple, à partir de la formule de Bayes, on obtient :

$$\log p(w_i|x) = \log p(x|w_1) + \log p(w_i) - \log p(x) \quad \textbf{(1.3)}$$

Il est alors possible d'utiliser $g_i(x) = \log p(x|w_i) + \log p(w_i)(i = 1,...,r)$ comme fonction discriminante. Spécialement dans le cas de deux catégories, on peut utiliser $g(x) = g_1(x) - g_2(x)$ comme fonction discriminante. On décide ensuite, que x appartient à w_1 si la condition $(g(x) > 0)$ est vérifiée sinon x appartient w_2 si $(g(x) < 0)$.

Pour appliquer la méthode de décision bayésienne, on doit connaître :

1. Les probabilités a priori $p(w_i)$ $(i = 1,\ldots,r)$.
2. Les densités de probabilité a posteriori $p(x|w_i)$ $(i = 1,\ldots,r)$.

Si nous supposons que $p(x|w_i)$ est une densité de probabilité gaussienne de n dimensions, alors $p(x|w_i)$ est donnée par :

$$p(x|w_i) = \frac{1}{\sqrt{(2\pi)^n |\sum_i|}} \exp\left\{-\frac{1_t}{2}(x - m_i)\sum_i^{-1}(x - m_i)\right\} + \log\, p(w_i) \quad \textbf{(1.4)}$$

Dans le cas particulier du problème de classification à deux catégories, une fonction discriminante de Bayes g est donnée par:

$$g(x) = {}^t x w_1 x - {}^t w2x + {}^t u1x - {}^t u2x + w_{10} - w_{20} \quad \textbf{(1.5)}$$

$$wi = -\frac{1}{2}\sum_i^{-1} \quad , u_i = \sum_i^{-1} m_i \quad \text{et} \quad w_{i0} = -\frac{1^t}{2} m_i \sum_i^{-1} m_i - \frac{1}{2}\log|\sum_i| + \log p(w_i)$$

Les Réseaux bayésiens

Le réseau bayésien est un modèle graphique permettant d'exprimer des relations de probabilité entre un ensemble de variables. La structure d'un réseau bayésien est considérée comme un graphe dirigé acyclique (GDA) et les nœuds sont constitués dans une correspondance *1-à-1* avec les caractéristiques X. Les arcs représentent des influences entre les caractéristiques, alors que l'absence d'un arc représente une indépendance conditionnelle. Un nœud est conditionnellement indépendant de ces non-descendants étant donnés ses parents. Par exemple : X1 est conditionnellement indépendant de X_2 étant donné X_3 si $P(X_1|X_2, X_3) = P(X_1,X_3)$ pour toutes les valeurs possibles de (X_1,X_2,X_3).

La caractéristique la plus importante des réseaux bayésiens par rapport aux arbres de décisions et aux réseaux de neurones, est certainement leur possibilité de prendre en compte des informations a priori pour un problème donné. Cette expertise préalable, ou la connaissance du domaine, sur la structure d'un réseau bayésien peut prendre les formes suivantes :

1. Déclarer qu'un nœud est une racine ou bien n'a pas de parents.
2. Déclarer qu'un nœud est une feuille ou bien n'a pas d'enfants.
3. Déclarer qu'un nœud est une cause directe ou un effet direct d'un autre nœud.
4. Déclarer qu'un nœud n'est pas directement connecté à un autre nœud.
5. Déclarer que deux nœuds sont indépendants, étant donné un ensemble de conditions.
6. Donner un ordre partiel des nœuds, c.à.d. un nœud apparaît plutôt qu'un autre nœud dans l'ordre.
7. Donner un arrangement de nœuds complet.

Le problème avec les réseaux bayésiens est qu'ils ne conviennent pas pour des bases contenant un trop grand ensemble de caractéristiques *(Cheng et al. [2002])*. La raison liée à ce phénomène est qu'il est difficile de construire un très grand réseau pour des raisons évidentes de temps de calcul et d'espace mémoire.

Le tableau 1.4 illustre les taux de reconnaissance de manuscrits par les réseaux bayésiens. Nous pouvons conclure que l'association caractéristiques-réseaux bayésiens est très importante et va avoir un effet sur les taux de reconnaissance. Les résultats de *(Wang et al. [2005])* et *(Sung et al.* [*2006])* confirment cette conclusion puisque sur le même type de documents la différence entre les taux de reconnaissance est de *17,3%.* Ce constat a été fait par Lambert Schomaker dans *(Schomaker, [2007])* qui a porté dans le cadre d'une aide à l'expertise judiciaire, une attention très scrupuleuse à la définition des descripteurs de formes écrites, de sorte qu'ils assurent une réelle exhaustivité et authorisent une véritable interprétation par les experts humains à qui l'analyse automatique de scripteurs se destine.

Tableau 1.4. Résultats des Réseaux bayésiens sur la classification des manuscrits

Auteur(s)	Caractéristiques	Type de document	Taux de reconnaissance
(Guo et al.[2010])	Ondelettes	chinois, anglais	*68,86 à 94,95%*
(Cho et al. [2003])	Code de chaîne de Freeman du contour	hangul	*95,7%*
AlKhateeb, [2012]	Nb. de pixels du premier plan, nb. De composantes connexes sous et en dessus de la ligne de base	arabe	*95,6%*
Wang et al. [2005])	Caractéristiques spatiales	anglais	*79,3%*
(Sung et al. [2006])	Filtres de Gabor	anglais (chiffres)	*96,6%*

3.1.4 Les méthodes basées sur les perceptrons

Le perceptron multicouche (PM)

Les perceptrons multicouches sont des classifieurs linéaires de type réseau de neurones organisés en plusieurs couches : ils permettent donc de classifier des ensembles d'instances qui sont linéairement séparables. Si une ligne droite ou un hyperplan peut être tracé pour séparer les instances en catégories, alors les instances sont linéairement séparables et un perceptron peut être construit. Si les instances ne sont pas linéairement séparables alors l'apprentissage ne va pas aboutir à classifier toutes les instances correctement. Les PM ont été utilisés pour résoudre les problèmes linéairement séparables. Un PM est formé à partir de neurones qui sont réunis selon une architecture de connections (figure 1.6). Les unités ou les neurones sont généralement séparés en trois classes :

1. Les unités d'entrée, qui reçoivent l'information à analyser.
2. Les unités de sortie, où les résultats de l'analyse sont trouvés.
3. Les unités cachées entre l'entrée et la sortie.

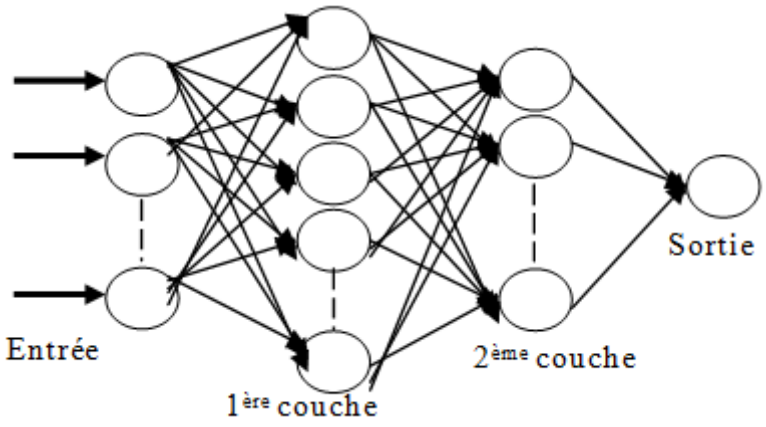


Figure 1.6. Modèle graphique du perceptron multicouche *(Dawson et Wilby, [2001])*

Le réseau de neurones « FeedForward » permet au signal de se propager dans une seule direction, de l'entrée vers la sortie (figure 6). Lors de la phase d'apprentissage, le réseau est entraîné sur un ensemble de données étiquetées pour organiser l'entrée et la sortie du perceptron, c'est à dire fixer les poids des connexions entre les neurones. Le réseau détermine alors la classification d'un nouvel ensemble de données. L'apprentissage se fait le plus souvent à partir de l'algorithme de rétro-propagation pour estimer les valeurs des poids. Le problème des algorithmes de rétro-propagation c'est qu'ils sont trop lents pour la plupart des applications, une approche pour contourner ce problème est l'estimation optimale des poids initiaux, ou l'utilisation des algorithmes génétiques pour l'apprentissage des poids des neurones du réseau.

Architecture des réseaux de neurones

L'apprentissage supervisé des réseaux de neurones peut être atteint à partir :

- De la modification de poids synaptiques.
- De la modification de l'architecture du réseau : créer ou supprimer les neurones ou bien des connections synaptiques.
- D'un choix approprié des fonctions d'activation comme par exemple la fonction logistique, la tangente hyperbolique, la fonction Gaussienne et la fonction à seuil.

Les réseaux de neurones à base radiale ont été utilisés dans un grand nombre de domaines d'ingénierie pour la résolution de problèmes de classification et de décision *(Howlett et Jain, [2001])*. Un réseau de neurones à base radiale est un réseau de rétroaction à trois couches dans lequel des unités cachées implémentent une fonction radiale d'activation et chaque unité de sortie implémente une somme pondérée des sorties des unités cachées. La procédure d'apprentissage est divisée en deux étapes. Tout d'abord les centres et les largeurs des unités cachées sont déterminés à partir d'algorithmes de clustering. Les poids reliant la couche cachée avec la couche de sortie peuvent ensuite être déterminés par l'algorithme de « décomposition en valeurs singulières (SVD)».

Malgré le fait que les réseaux de neurones et les arbres de décision sont des techniques différentes, quelques chercheurs comme (*Eklund et Hoang, [2002]), (Tjen-Sien Lim et al. [2000])* ont conduit des études comparatives empiriques. Et voici quelques conclusions :

- Les réseaux de neurones sont généralement plus en mesure de fournir un apprentissage incrémental que les arbres de décision.
- Le temps d'apprentissage pour les réseaux de neurones est plus long que le temps d'apprentissage pour les arbres de décision.
- Les réseaux de neurones fournissent des performances souvent bien meilleures que celles produites par les arbres de décision.

Les réseaux de neurones à base radiale ont de nombreux avantages : ils montrent une bonne résistance au bruit et au manque de fiabilité des données, et sont capables de passer directement des données au prédicateur, sans intermédiaire, sans recodage, sans discrétisation, et sans simplification sujets à caution. Ils ont une capacité à représenter n'importe quelle dépendance fonctionnelle.

Les réseaux de neurones ont été appliqués à de nombreux problèmes du monde réel. Leur point faible réside dans leur incapacité à justifier leur sortie d'une manière compréhensible pour l'utilisateur. Le phénomène de « boîte noire » reste un inconvénient important pour

l'interprétation des résultats. Cela est la principale différence entre les réseaux de neurones et les arbres de décision. Si l'utilisateur a besoin de pouvoir interpréter le résultat de l'apprentissage, il choisira préférentiellement un système basé sur les arbres de décision, sinon les deux méthodes sont concurrentes.

Le tableau 1.5 illustre les taux de reconnaissance atteints par les réseaux de neurones pour la classification des manuscrits. Ici aussi le taux de reconnaissance va varier selon le type de caractéristiques utilisées. Pourtant la différence de reconnaissance entre *(Roy et al. [2004])* et *(Patil et Subbareddy, [2002])* n'est pas très importante, mais nous pouvons aussi conclure que les caractéristiques affectent les résultats.

Tableau 1.5. Résultats des Réseaux de neurones sur la classification des manuscrits

Auteur(s)	**Caractéristiques**	**Type de document**	**Taux de reconnaissance**
(Elgammal et. Ismail. [2001])	Proj. Horizontale, Moments, Sommets, Distribution Run-Length	arabe, latin	*96,8%*
(Patil et Subbareddy, [2002])	Nombre de pixels	anglais, indien, kannada	*96%*
(Roy et al. [2004])	Fractales, Topologiques…	indien	*97,62%*
(Hangarge et Dhandra, [2009])	Longueur du trait (Nombre de pixels), Densité des pixels	anglais, devanagari, urdu	*88,6% à 99,2%*

3.2 Les méthodes de classification non-supervisées

Dans cette section nous présentons les méthodes de classification non-supervisées avec leurs avantages, inconvénients et les applications dans le domaine de classification de manuscrit.

3.2.1 Les méthodes incrémentales

La classification incrémentale non-supervisée est basée sur le fait qu'il est possible de considérer les formes une à la fois et de les affecter à des groupes existants. L'algorithme attribue le premier élément à un groupe, l'élément suivant est soit affecté à l'un des clusters existants ou bien à un nouveau groupe qui est créé de façon incrémentale. Cette affectation se fait en se basant sur un critère comme la distance entre ce nouvel élément et le centroïde du groupe existant. Le même principe est appliqué au reste des éléments. Parmi ces méthodes nous distinguons :

L'algorithme *Leader*

Cet algorithme a été proposé par *(Hartigan, [1975]).* Il ne nécessite pas de connaître a priori le nombre de classes désiré et fonctionne en un seul passage, ce qui est important durant la classification des grandes bases de données. Dans *(Isaacman et al. [2011]),* les auteurs ont utilisé l'algorithme *Leader* pour identifier les emplacements généralement importants et

discerner les endroits sémantiquement significatifs tels que la maison et le travail sur une base de données de *4 GB*.

L'algorithme du plus court chemin couvrant

Cet algorithme a été proposé par Slagle dans *(Slagle et al. [1975])* pour la réorganisation de données et a été utilisé avec succès dans l'audit automatique des enregistrements *(Lee et al. [1978])*. Dans ces travaux l'algorithme est utilisé pour regrouper 2000 formes en utilisant 18 caractéristiques. Ces groupes sont utilisés pour estimer les caractéristiques manquantes.

L'algorithme cobweb

L'algorithme cobweb proposé par *(Fisher, [1987])*, est une approche de classification conceptuelle, basée sur les notions de *category utility(CU)* et *partition utility (PU)*. Cet algorithme est appliqué sur des variables qualitatives, et utilise un apprentissage incrémental. Au lieu de suivre une approche de classification divisive ou agglomérative, il construit dynamiquement un dendrogramme en passant en revue les individus un à un. Dans cobweb chaque classe est considérée comme un modèle plutôt que comme une collection d'individus qui la compose, c'est pour cela qu'il est considéré comme une approche conceptuelle. Chaque segment d'arbre est associé à des probabilités conditionnelles de modalités des variables $P(V_j = m_j | C_r)$. Au cours de la construction du dendrogramme, chaque nouvel individu parcourt l'arbre qui est au fur et à mesure mis à jour en utilisant les valeurs de *CU* et *PU*.

L'algorithme de groupement incrémental pour le traitement dynamique d'informations

Un algorithme de regroupement incrémental pour le traitement de l'information dynamique a été présenté par Can dans *(Can, [1993])*. La motivation qui soutient ce travail est que, dans les bases de données dynamiques, des éléments peuvent être ajoutés et supprimés au fil du temps.

L'avantage des méthodes incrémentales c'est qu'il n'est pas nécessaire de stocker toute la matrice de données en mémoire. Ainsi, les besoins en espace des algorithmes incrémentaux sont très petits. Comme ils ne sont pas itératifs, le temps de calcul est également réduit *(Jain et al. [1999])*. Cependant cet algorithme peut engendrer des difficultés lors de l'insertion de nouveaux éléments, ne garantissant pas toujours l'exactitude des résultats *(Shaw et Yu, [2009])*.

Le tableau 1.6 illustre les taux de reconnaissances de systèmes basés sur des méthodes incrémentales. Dans *(Bensefia et al. [2005b])* et *(Siddiqi, [2009])*, les auteurs ont testé leur méthode d'identitification de scripteur sur la base IAM mais avec des ensembles différents de caractéristiques. Les résultats montrent que les caractéristiques utilisées par *(Bensefia et al. [2005b])* permettent d'avoir de meilleurs taux de reconnaissances, ce qui montre l'importance du choix des caractéristiques et leurs impacts dans les méthodes d'identification de scripteurs et de reconnaisance de style.

Tableau 1.6. Résultats des méthodes incrémentatles pour l'identification de scripteurs

Auteur(s)	Caractéristiques	Types de document	Taux de reconnaissance
(Bensefia et al. [2003])	Graphèmes	francais	*93,3%*
(Nosary et al. [2004])	Caractéristiques graphiques des scripteurs	francais	*≈96%*
(Bensefia et al. [2005b])	Graphèmes	anglais	*86%-100%*
(Siddiqi, [2009])	Caractéristiques géométriques	anglais	*91%*

3.2.2 Les méthodes de type partition

Les nuées dynamiques

Dans cette catégorie de méthodes, les classes sont représentées par un noyau. Ce noyau peut être représenté soit comme le centre de gravité d'une classe (comme cela est le cas dans l'approche des centres mobiles ou *k*-moyennes), soit par un ensemble d'individus comme dans l'approche des k-médianes), soit par une distance comme dans l'approche adaptative de Diday *(Diday et Simon, [1976]),* soit enfin par une loi de probabilité comme c'est le cas dans les travaux de *Schroeder (Schroeder, [1976]).*

Le principe de cet algorithme est le suivant :

Considérons un ensemble d'individus appartenant à un ensemble *E*. L'objectif est de chercher la meilleure partition à *k* classes fixées de cet ensemble selon le critère d'inertie.

Cet algorithme est itératif et à chaque étape la qualité de la partition s'améliore. Le nombre de classes est connu a priori, ainsi que le nombre d'éléments au centre du noyau qui seront énumérés. Au début l'ensemble des noyaux d'une classe est sélectionné au hasard. Les éléments les plus proches de ces noyaux sont regroupés autour de ces derniers, et la distance euclidienne est utilisée pour calculer la distance par rapport au centre de la classe. A partir de cette partition, une autre famille de noyaux est déterminée, et elle rassemble les points les plus proches formant ainsi une nouvelle classe. Ce processus itératif continu jusqu'à atteindre un nombre fini de classes. Si après un certain nombre d'itérations, on a des classes stables, alors les données sont dites classifiables et constituent des formes fortes (figure 1.7).

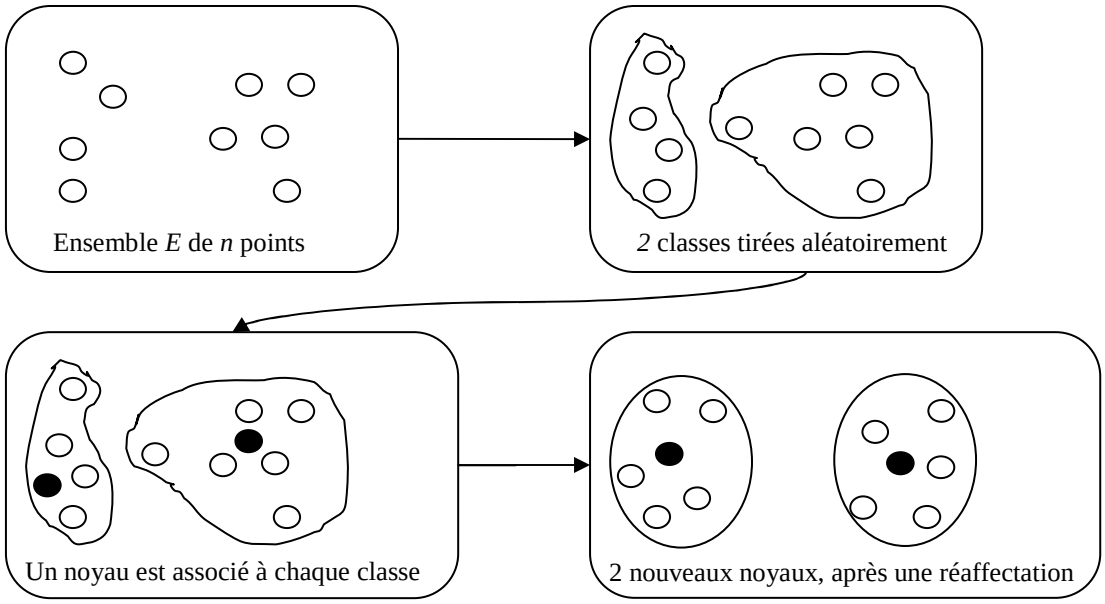


Figure 1.7. Principe de la méthode des nuées dynamiques avec ($k \leq 2$), *(Diday, [1971])*

Soit un ensemble E de n points, nous cherchons à créer une partition de E en k classes. Chaque classe est présentée par son noyau. Une bonne classification est obtenue si et seulement si la somme des distances entre les individus et les noyaux est minimale.

Nous détaillons deux méthodes les plus classiques qui relèvent de ce principe.

La méthode des *k*-moyennes

Cette méthode consiste à construire une partition en k classes en sélectionnant k individus comme centres des classes, ils sont tirés au hasard dans l'ensemble des individus. Après cette sélection, chaque individu est associé au centre le plus proche ce qui donne une partition en k classes, les centres des classes seront actualisés et de nouvelles classes seront formées suivant le même principe. L'entier k désigne le nombre maximum de classes désiré. Généralement la partition obtenue est localement optimale car elle dépend du choix initial des centres. Pour cela les résultats entre deux exécutions de l'algorithme peuvent varier de façon significative. Cette méthode est simple, compréhensible et est applicable à des données dans un espace de grande dimension. La contrainte principale est que le nombre de classes doit être fixé au départ.

La méthode des *k*-médoïdes

Dans les méthodes des k-médoïdes une classe est représentée par un de ses individus (médoïde). C'est une méthode itérative combinant la réaffectation des individus dans des classes avec une intervention des médoïdes et des autres individus. C'est une méthode simple et elle couvre n'importe quel type de variables. Quand les méthodes des k-médoïdes sont choisies, des classes sont définies comme des sous-ensembles d'individus proches des médoïdes les plus proches par rapport à distance choisie. Le médoïde d'un groupe est l'individu possédant la dissimilarité moyenne des individus du groupe. La méthode des *k*-médoïdes est considérée comme une version généralisée de la méthode des k-moyennes, cette dernière peut seulement

être appliquée avec des variables numériques et quantitatives. Les médoïdes se calculent également sur des données catégorielles, et l'algorithme des k-médoïdes nécessite des capacités calculatoires plus importantes que celui des k-moyennes. Cette méthode est flexible autorisant l'usage de plusieurs types de distances, mais elle nécessite également de spécifier a priori le nombre de classes *k*.

Le tableau 1.7 illustre les taux de reconnaissance de systèmes basés sur des méthodes *k*-moyennes et *k*-médoïdes. Ce qui est intéressant de noter à partir de ces résultats, c'est l'importance de la combinaison des caractéristiques, bien plus efficace qu'un simple usage individuel de chacune d'elles pour classifier les manuscrits. Bulacu dans *(Bulacu et Schomaker, [2007])* montrent en quoi la combinaison de caractéristiques améliore les taux de reconnaissance.

Tableau 1.7. Résultats des méthodes *k*-moyennes, *k*-médoïdes et nuées dynamiques sur la classification des manuscrits

Auteur(s)	Caractéristiques	Types de document	Taux de reconnaissance
(Schomaker et al. [2007])	Caractéristiques basées sur le contour	hollandais	*64,7% à 92,7%*
(Bulacu et Schomaker, [2007])	Textures, Allographes	arabe	*19% à 97% (caractéristiques individuelles), 38% à 99%(combinaison de caractéristiques)*
(Nguyen et al. [2008])	Différences de Gaussiennes	symboles	*90%*

3.2.3 Les méthodes de type hiérarchique

La construction d'une classification hiérarchique peut se faire de deux façons : pour la première, à partir d'une matrice symétrique des similarités entre les individus, un algorithme agglomératif forme initialement de petites classes ne comprenant que des individus très semblables, puis, à partir de celles-ci, il construit des classes de moins en moins homogènes, jusqu'à obtenir une unique classe. Ce mode de construction est appelé Classification Ascendante Hiérarchique (CAH). Le second mode de construction d'une classification hiérarchique inverse le processus précédent. Il repose sur un algorithme divisif muni d'un critère de division d'un sous-ensemble de variables, et procède par dichotomies successives de l'ensemble des individus tout entier, jusqu'à un niveau vérifiant certaines règles d'arrêt et dont les éléments constituent une partition de l'ensemble des individus à classer. Ce mode de construction s'appelle la classification.

3.2.3.1 Classification Ascendante Hiérarchique (CAH)

Classification Ascendante Hiérarchique (CAH)

Le but d'une CAH est d'obtenir une classification automatique de l'ensemble des individus. Cette classification commence par déterminer parmi les n individus, quels sont les *2* individus qui se ressemblent le plus par rapport à l'ensemble des p caractéristiques spécifiées. Elle va alors regrouper ces deux individus pour former une classe. Il existe donc à ce niveau $n-1$ classes, une étant formée des deux individus regroupés précédemment, les autres ne contenant qu'un unique individu. Le processus se poursuit en déterminant quelles sont les deux classes qui se ressemblent le plus, et en les regroupant. Cette opération est répétée jusqu'à l'obtention d'une unique classe regroupant l'ensemble des individus.

Cette procédure est basée sur la détermination d'un critère de ressemblance entre les individus. La méthode laisse à l'utilisateur le choix de la dissimilarité et la détermination d'une dissimilarité entre classes : procédé appelé un critère d'agrégation. Le critère d'agrégation permet de comparer les classes deux à deux pour sélectionner les classes les plus similaires. Les critères les plus classiques sont le plus proche voisin, le diamètre maximum, la distance moyenne et la distance entre les centres de gravité.

Classification symbolique ascendante hiérarchique

En 2003, une approche symbolique de classification ascendante hiérarchique a été proposée par *(Mali et Mitra, [2003]).* Elle suit le même principe de fonctionnement que les approches classiques mais en diffère par le critère d'agrégation qu'elle utilise. En effet, elle définit la distance entre deux classes C_1 et C_2 comme suit :

$$d_{agrégation}\ (C_1, C_2) = \frac{\sum\sum d(x_u, x_q)}{|C_1||C_2|}\left(\frac{|C_1||C_2|}{|C_1|+|C_2|}\right)^{\frac{1}{2}} \qquad \textbf{(1.6)}$$

où d représente la mesure de dissimilarité de Gowda et Diday *(Gowda et Diday, [1992])* définie sur l'ensemble des individus et |Ci| représente le cardinal de la classe Ci.

On observera notamment que le terme de pondération utilisé par cette distance prend une valeur de $\sqrt{50}/10000$ pour *($|c_i| = |c_j| = 100$),* une valeur de $\sqrt{1/10100}$ pour *($|c_i| = 1$ et $|c_j| = 100$)* et une valeur de $\sqrt{0{,}5}$ pour *($|c_i| = |c_j| = 1$).* En conséquence, l'approche de classification hiérarchique tend à favoriser le fusionnement des classes singletons, ou des petites et grandes classes, au détriment de la fusion des classes de tailles moyennes.

Ces méthodes de classification ascendante hiérarchique sont faciles à implémenter. Mais elles sont très coûteuses avec une complexité temporelle en $O(n^2)$.

3.2.3.2 Classification Descendante Hiérarchique (CDH)

Les méthodes de classification descendante hiérarchique partent d'un ensemble d'individus et construisent, de manière itérative, une partition de l'ensemble. A l'inverse de la classification ascendante hiérarchique, à chaque étape, l'algorithme se charge de deux actions :

1. Chercher une classe à diviser
2. Choisir un mode d'affectation des objets aux sous-classes

Parmi les algorithmes les plus anciens, l'algorithme de Williams présentés dans *(Williams et Lambert, [1959])* divise la classe la plus grande en deux classes. Hubert dans *(Hubert, [1973])* a proposé de diviser la classe de plus grand diamètre. Aucun des deux n'a justifié son choix de division.

Cette méthode de classification construit une hiérarchie, en commençant par une grande classe contenant tous les objets. A chaque étape, elle divise une classe en deux classes plus petites jusqu'à ce que toutes les classes ne contiennent qu'un seul individu. Ceci veut dire que pour n individus, la hiérarchie est construite en $n-1$ étapes au plus. Dans la première étape, les données sont divisées en deux classes au moyen des dissimilarités. Dans chacune des étapes suivantes, la classe avec le diamètre le plus grand se divise de la même façon. Après $n-1$ divisions, tous les individus sont bien séparés. La dissimilarité moyenne entre l'individu x qui appartient à la classe C contenant n individus et tous les autres individus de la classe C est définie par :

$$d_x = \frac{1}{n-1} \sum_{x \in C,\, y \neq x} d(x,\, y) \qquad \textbf{(1.7)}$$

Par rapport à la plupart des algorithmes en classification automatique, l'algorithme de classification descendante hiérarchique ne nécessite pas l'utilisation d'un seuil arbitraire pour la formation des classes qui peut éventuellement mener à une partition non réaliste. Si l'algorithme d'échange ne privilégie pas les aspects locaux, il est initialisé avec une partition liée par des relations de filiation avec des partitions précédemment obtenues. Cela donne à l'algorithme un certain aspect global. Les résultats sont en général grossiers, les niveaux des nœuds de la hiérarchie ne sont plus définis par l'ordre dans lequel ils apparaissent.

Bien que les méthodes hiérarchiques représentent la famille principale des techniques de classification et qu'elles aient été appliquées avec succès dans plusieurs domaines, elles

souffrent d'une faiblesse qui réside dans leur critère de partitionnement qui n'est pas global, mais dépend des classes déjà obtenues précédemment. En effet, les opérations de fusions/divisions des classes se déroulent sans jamais remettre en cause les associations déjà constituées, ce qui peut conduire à des classes peu représentatives (notamment en présence de données aberrantes) *(NG et Han, [2002])*. Pour les cas agglomératifs par exemple, deux individus placés dans des classes différentes ne sont plus jamais comparés, et deux individus placés dans une même classe ne peuvent plus être séparés. En d'autres termes, la classification obtenue en *k* classes est rarement la meilleure possible (optimale), mais seulement la meilleure entre celles obtenues en fusionnant des classes d'une classification en *k+1* classes.

Le tableau 1.8 illustre les taux de reconnaissance obtenus par des méthodes hiérarchiques. Ces méthodes permettent d'avoir de bons taux de reconnaissance de manuscrits, mais ces résultats ne sont pas suffisants pour conclure à leur réelle performance, car il est nécessaire de tenir compte du choix des caractéristiques sur les taux de reconnaissance.

Tableau 1.8. Résultats des méthodes de classification hiérarchique sur la classification des manuscrits

Auteur(s)	Caractéristiques	Types de document	Taux de reconnaissance
(Hochberg et al. [1997])	Nb. de pixels noirs et blancs	arabe, cyrillique, grecque, hébreux…	*98%*
(Bhardwaj et al. [2009])	Orientation, courbure à partir du contour	anglais- base IAM	*75,5%*

3.2.4 Autres approches particulières de classification

Dans chacune des approches de classification automatique que nous venons de détailler ci-dessus, les méthodes sont fondées sur la notion de distance. D'autres méthodes de classification ont été développées, elle peuvent être divisées en deux groupes selon la définition donnée à la notion de classe. Nous trouverons ainsi les méthodes fondées sur la notion de densité et les approches fondées sur un modèle.

3.2.4.1 Les approches fondées sur la notion de densité

Les approches fondées sur la densité supposent que les points qui appartiennent à chaque groupe sont tirés d'une distribution de probabilité spécifique *(Banfield et Raftery, [1993])*. La répartition globale des données est supposée être un mélange de plusieurs distributions. Parmi les approches fondées sur la densité citons les algorithmes suivants :

L'algorithme DBSCAN trouve des groupes de formes arbitraires et est efficace pour les grandes bases. L'algorithme constitue des groupes en cherchant dans le voisinage de chaque

élément de la base s'il contient plus que le nombre minimum d'objets *(Ester et al. [1996]).* L'algorithme AutoClass est un algorithme largement utilisé qui couvre une grande variété de distributions, y compris gaussienne, de Bernoulli, de Poisson, et distributions log-normales *(Cheeseman et Stutz, [1996]).*

D'autres méthodes bien connues à base de densité incluent: SNOB *(Wallace et Dowe, [1994])* et MCLUST *(Fraley et Raftery, [1998]).* Les approches fondées sur la notion de densité présentent l'intérêt de trouver elles-mêmes une évaluation du nombre de classes, l'algorithme de maximisation de l'attente 'EM' *(Dempster et al. [1977])* permet d'estimer ces paramètres. Ces approches acceptent tout type de données et prennent en compte les données aberrantes qui ne sont pas affectées aux groupes.

Mais les expériences montrent que les résultats obtenus sont très sensibles aux choix des paramètres *(Nakache et Confais, [2005]).* Nous n'avons pas trouvé de méthodes d'identification de scripteurs ou de classification de manuscrits dans la littérature reposant sur ces approches de classification. En revanche cette méthode a été utilisée pour séparer les annotations manuscrites des textes imprimés *(Mazzei et al. [2009])* et pour séparer le texte des images *(Liu et al. [2008]).*

3.2.4.2 Les cartes auto-organisatrices (SOM)

Les cartes auto-organisatrices ont été introduites par *(Kohonen, [1982])* et sont initialement appliquées à l'image et au son. Elles représentent aussi un mécanisme efficace de classification pour les données alphanumériques et font partie de la famille des réseaux de neurones.

Une carte auto-organisatrice est un procédé qui convertit un signal d'entrée complexe (plusieurs variables par exemple) en une nouvelle variable catégorielle : c'est donc un procédé de classification (modélisation non-supervisée). Les *SOM* sont une généralisation de l'analyse en composantes principales. Elles fonctionnent comme un réseau de neurones sans variable cible et avec plusieurs nœuds dans la couche de sortie. La carte structure les nœuds en sortie en classes de nœuds (figure 1.8).

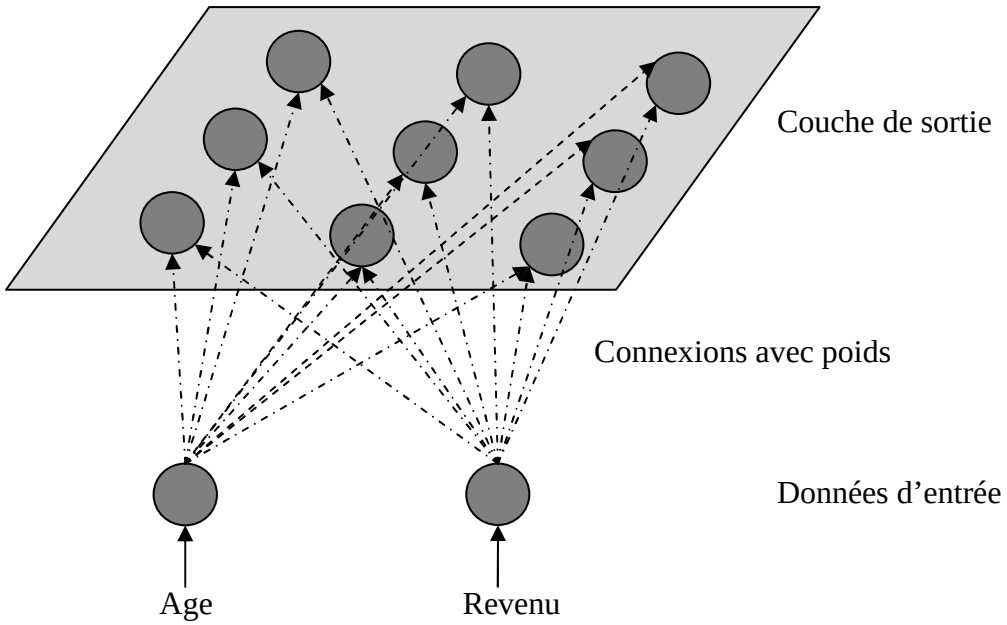


Figure 1.8. Schéma d'une carte auto-organisatrice *(Kohonen, [1982])*

A la différence des réseaux de neurones, les *SOM* n'ont pas de couche cachée. La couche de sortie contient plusieurs nœuds représentés sous la forme d'un treillis rectangulaire (un carré de dimension 3 dans l'exemple ci-dessus). Le nombre de nœuds de la couche de sortie est fixé arbitrairement par l'utilisateur qui définit le nombre maximum de classes.

Principe de fonctionnement : fonction de score

Les valeurs des nœuds de la couche d'entrée (valeurs normalisées des variables prises en compte par le modèle) sont distribuées dans les nœuds de la couche de sortie après transformation en fonction des pondérations du réseau : nous parlons de ' fonction de score'. Cette fonction est généralement une fonction de distance euclidienne entre les poids associés aux connections et les valeurs associées aux données d'entrée. Le nœud de sortie qui a le meilleur résultat 'meilleur score' est le 'nœud gagnant' : il reçoit l'individu en question. Le meilleur score, c'est la plus petite distance entre les poids de connexion et les données d'entrée.

Principe de fonctionnement : liaison de voisinage des nœuds de la couche de sortie

Les nœuds d'une même couche, et particulièrement de la couche de sortie, ne sont pas reliés entre eux. Toutefois, les poids des nœuds de voisinage du nœud gagnant sont adaptés pour favoriser leur victoire en cas de données similaires. C'est ce qu'on appelle la coopération et l'adaptation des nœuds de la couche de sortie. L'adaptation, c'est ce qui correspond à l'apprentissage.

Les réseaux de Kohonen sont des cartes auto-organisatrices avec une variation dans la technique d'apprentissage. Dans les réseaux de Kohonen, les nœuds dans le voisinage du nœud

gagnant ajustent leurs poids en utilisant une combinaison linéaire du vecteur d'entrée et du vecteur de poids en cours.

Comme avantages, notons que ces méthodes offrent une simplicité de calcul, et permettent d'avoir une classification raffinée. Par contre la lecture des résultats n'est pas facile, les classes sont très déséquilibrées et dépendent fortement de l'initialisation puisque le nombre de classes maximum est fixé a priori *(Fort et al. [2002]).*

Le tableau 1.9 illustre les taux de reconnaissance de manuscrits en utilisant les cartes auto-organisatrices. A partir de ce tableau nous pouvons conclure que la combinaison des caractéristiques affecte les taux de reconnaissance *(Bulacu et schomaker, [2006])* comme c'est le cas avec les méthodes k-moyennes et k-médoïdes.

Tableau 1.9. Résultats des cartes auto-organisatrices sur la classification des manuscrits

Auteur(s)	Caractéristiques	Types de document	Taux de reconnaissance
(Bulacu et Schomaker, [2005])	Caractéristiques basées sur le contour	arabe,anglais	*92,2% (Ksom 1D) , 92,6% (Ksom 2D)*
(Bulacu et Schomaker, [2006])	Textures, Allographes	arabe	*97% (caractéristiques individuelles),99%(combinaison de caractéristiques)*
(Moghaddam et Cheriet, [2009])	Niveau de gris du trait, Niveau de gris des pixels du fond	arabe	*92,6% à 94,8%*
(Schomaker et al. [2007])	Composantes connexes	hollandais	*99.7%*

3.3 Méthodes de classification semi-supervisées

Les méthodes de classification semi-supervisées utilisent en même temps les données étiquetées et non-étiquetées pour classifier les données. Nous distinguons en général deux types d'approches utilisés par les méthodes de classification semi-supervisées :

- La première consiste à regrouper les éléments les plus similaires en se basant sur l'information apportée par les données étiquetées. A l'issue de l'apprentissage, deux éléments qui ont des étiquettes différentes ne peuvent pas appartenir au même groupe. Parmi ces approches notons les travaux de Cohn dans *(Cohn et al. [2003]).* Cette approche considère des informations supplémentaires qui permettent de spécifier dans l'ensemble d'apprentissage, indépendamment des étiquettes, si deux éléments doivent ou non appartenir à un même groupe.
- La deuxième approche proposée par *(Chapelle et Zien, [2005])* utilise en premier les données étiquetées pour séparer les éléments en fonction de leurs étiquettes. Puis, les données non-étiquetées sont utilisées pour affirmer le modèle obtenu. Dans *(Chapelle,*

[2005]) les données étiquetées sont utilisées pour estimer les densités de probabilité de chacun des groupes.

Le tableau 1.10 illustre des cas où l'apprentissage semi supervisé a été utilisé dans le cas d'identification de scripteurs et de classification de manuscrits.

Tableau 1.10. Résultats des méthodes semi-supervisées pour la classification de manuscrits et l'identification de scripteurs

Auteur(s)	Caractéristiques	Types de document	Taux de reconnaissance
(Ball et Srihari, [2009])	Combinaison de caractéristiques locales et globales	anglais, arabe	*86% et 77% pour la méthode semi-supervisée contre 81% et 76% pour la méthode supervisée*
(Kourtis et Stamatatos, [2011])	n-gram	courriels (textes de petites et grandes tailles)	*63,8% pour les textes de petite taille et 65,8% pour les textes de grande taille.*

3.4 Bilan sur les méthodes de classification supervisées, non-supervisées et semi-supervisées

Nous avons présenté dans cet état de l'art, les méthodes de classifications supervisées, non-supervisées et semi-supervisées qui ont été utilisées pour la classification de manuscrits et l'identification de scripteurs. Ce qui est important à noter c'est l'impact des caractéristiques sur le taux de reconnaissance indépendamment du classifieur. Bien sûr le choix du classifieur est important, dans nos travaux il faut prendre en considération que nous ne possédons pas de vérité de terrain à utiliser durant l'étape de classification, donc les algorithmes supervisés ne peuvent pas être utilisés puisque ni le nombre de classes ni l'étiquetage ne sont connus.

Aussi les méthodes semi-supervisées sont liées à la connaissance du nombre de classes. Comme nous n'avons pas de connaissance a priori de l'étiquette associée aux écritures, nous ne connaissons pas le nombre de classes. Par conséquent l'utilisation des méthodes de classification non-supervisées est indispensable, et comme nous avons remarqué, il existe beaucoup de méthodes non-supervisées. Il nous appartient donc de déterminer quelle méthode est la plus adaptée à notre problématique.

Dans la section suivante, nous proposons une méthode de classification non-supervisée basée sur la coloration de graphe. Elle n'a jamais été utilisée dans le domaine de la classification de manuscrits médiévaux mais elle a montré sa pertinence dans de nombreux domaines comme celui de la classification de courriers postaux.

4 Choix de la méthode de coloration de graphe

Les méthodes de partitionnement par coloration de graphe ont souvent été utilisées pour résoudre certains problèmes de planification (organisation de planning d'examens en particulier). La littérature propose deux grandes variétés de méthodes :

- Les méthodes directes : fondées sur la programmation dynamique *(Wang et Christofides, [2008])* ou sur des procédures d'énumération du type « Branch and Bound » proposée par *(Land et Doig, [1960]),* ces méthodes permettent de déterminer la valeur exacte du nombre chromatique en exigeant un temps de calcul défini comme une fonction non polynomiale du nombre de sommets. Le problème de la coloration d'un graphe est en effet *NP* complet.
- Les méthodes approchées : moins gourmandes en temps de calcul, elles s'appuient sur des considérations heuristiques. La qualité des estimations du nombre chromatique qu'elles donnent dépend de la structure du graphe *(Matula et al. [1972]).*

Nous proposons ici une méthode d'un esprit différent en considérant le problème de la coloration d'un graphe pour résoudre un problème de classification.

Nous nous sommes intéressé ici à la coloration de graphe dans un sens d'optimisation d'un critère lié à la finesse du partitionnement produit. C'est un mécanisme itératif d'optimisation qui est engagé. La dépendance de ce critère à l'égard du nombre de classes n'est pas univoque : nous pouvons ainsi dire que le problème de classification lui-même est un problème approché mais la méthode employée est une méthode exacte. Elle peut ainsi s'interpréter de deux façons : comme une méthode d'optimisation (d'un critère que nous devons définir basé sur des considérations de variances intra et inter classes) mais également comme une méthode de classification hiérarchique ascendante *(Benzécri et Benzécri, [1980]), (Jambu et Lebeaux, [1978]).*

Nous souhaitons dans notre travail adapter la coloration de graphe à un mécanisme de classification non-supervisé. Le problème se ramène alors à la détermination du nombre minimal de classes nécessaires pour diviser un ensemble de n objets en plusieurs sous-ensembles cohérents dont on cherchera à optimiser un critère. Il suffit donc de présenter chaque objet i par un sommet v_i et d'ajouter une arête $E(v_i, v_j)$, entre chaque paire d'objets pour lesquels on peut assurer qu'ils ne peuvent appartenir au même regroupement. Le graphe $G = (V, E)$ est défini par l'ensemble fini $V = \{v_1, v_2,\ldots,v_n\}$ $(|V| = n)$ dont les éléments sont appelés sommets, et par l'ensemble fini $E = \{e_1,e_2,\ldots,e_m\}$ $(|E| = m)$, dont les éléments sont appelés arêtes.

La coloration : La coloration d'un Graphe *G* est une fonction qui fait correspondre à chaque sommet *V* du graphe *G* un entier représentant une couleur dans $[1,...,|V|]$, en respectant la propriété que deux sommets adjacents n'ont pas la même couleur et que ces couleurs vont correspondre aux différentes classes auxquelles les sommets vont appartenir. Le problème d'optimisation revient alors à trouver une coloration avec un nombre de couleurs minimum.

Le nombre chromatique d'un graphe G (noté $\chi(G)$) est le nombre minimum de couleurs nécessaires pour lui attribuer une coloration propre.

Soit $\Omega = \{\omega_1,...,\omega_n\}$ l'ensemble de n éléments à classifier et d une mesure de dissimilarité sur Ω . On définit un graphe *G =(V, E)* sur Ω, associant à chaque sommet de *V* un élément de Ω. Une arête relie deux sommets du graphe si la dissimilarité entre les éléments correspondants de Ω est supérieure à la valeur d'un paramètre de contrôle α. Dans *(Bordat, [1986])* un algorithme de classification non-hiérarchique basé sur la coloration de graphe a été proposé : il exploite un indice de classification permettant d'identifier la partition qui s'ajuste au mieux à la structure de l'ensemble donné. Cet indice compte le nombre d'arêtes manquantes pour que chaque paire de sommets appartenant à des classes différentes soient adjacents, ce qui correspond à la situation où la dissimilarité entre éléments de classes différentes est toujours supérieure à α, et entre éléments d'une même classe toujours inférieure à α. Le nombre de classes et la partition qui s'ajuste au mieux aux données (parmi celles obtenues pour différentes valeurs de α), sont identifiés par un minimum local des valeurs de l'indice de classification.

Nous constatons que cette technique possède un coût de calcul raisonnable, avec un nombre minimal de contraintes. Dans le cadre de classification de lettres postales *(Gaceb et al. [2007])* les contraintes rencontrées étaient assez nombreuses. La classification par coloration a fourni des résultats très intéressants pour la localisation du bloc adresse et la classification de documents par le contenu. Les contraintes intégrées au système de coloration étaient alors les suivantes :

- La très grande variété de documents (de structures variées avec des contenus textuels manuscrits ou imprimés, sur des supports papiers dont la qualité, les couleurs et la texture peuvent être très différentes).
- Le fonctionnement en temps réel (quelques fractions de secondes doivent suffire à la reconnaissance).
- La maîtrise de la qualité des résultats (le système doit être le plus performant possible pour éviter le coûteux traitement manuel).

- Le type de document doit être identifié automatiquement malgré les aléas de numérisation (rotations, décalages, plissements du papier).
- La résolution spatiale des images élevée (*300* dpi).
- La superposition de couches d'informations (tampons, notes manuscrites, …).

4.1 Flexibilité de la méthode de coloration de graphe

Dans les techniques de classification hiérarchique, il est nécessaire de procéder à une coupure de l'arbre de classification (dendrogramme) par une droite horizontale qui fournira alors une partition possible de l'ensemble des individus à classer. Le nombre de classes de la partition est défini par le niveau de cette coupure, qui n'est pas toujours facile à déterminer : les partitions résultantes pourront ensuite être comparées afin de retenir la meilleure en termes de compacité et de séparabilité des classes.

Pour les techniques de classification par partitionnement, comme cela est le cas avec la coloration, si le nombre de classes à découvrir n'est pas fixé a priori, c'est une optimisation de la qualité des partitions résultantes qui conditionnera ce nombre basé sur un seuil unique de dissimilarité entre vecteurs descripteurs des instances. La flexibilité de la méthode de coloration de graphe permet de l'appliquer à n'importe quel niveau de classification : de la plus simple entité (qui représente un graphème) au dictionnaire de formes (pour l'exploitation complète des pages de manuscrits) (figure 1.9). Dans notre cas, nous avons exclusivement exploité la coloration de graphe pour la construction des dictionnaires de formes, l'analyse des styles d'écritures étant traité par application CBIR et par l'estimation de distances entre dictionnaires sans passage par un processus complet de clustering.

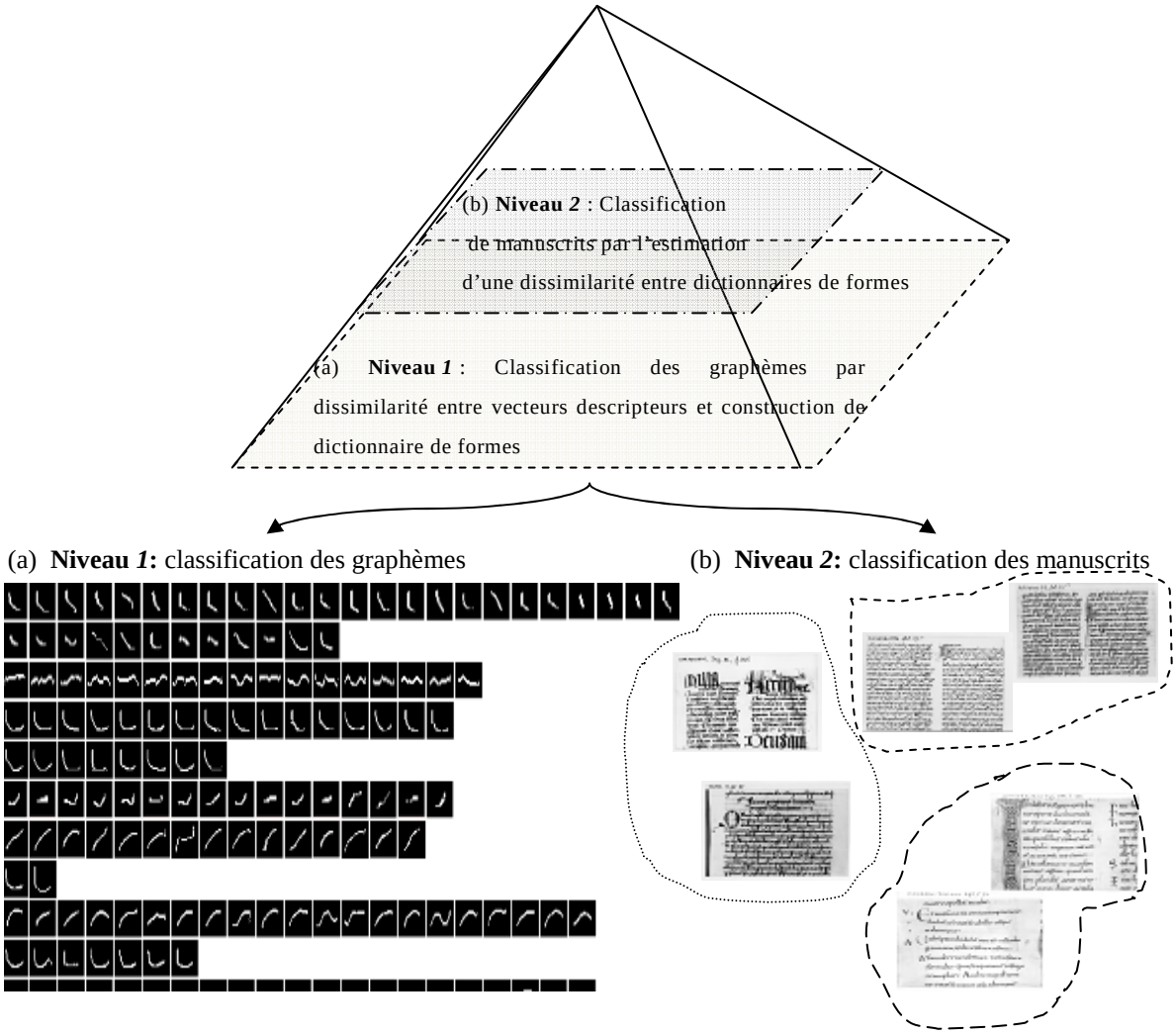


Figure 1.9. E**xploitation de la coloration de graphe (a) pour la construction des dictionnaires de formes, (b) pour la classification des manuscrits à partir d'un clustering des dictionnaires de formes**

4.2 Comparaison de la méthode de coloration de graphe avec les autres méthodes

La technique de coloration de graphe proposée par *(Gaceb et al. [2009])* pour le classement de courriers a été comparée avec les k-moyennes et les *SVM*. Les résultats montrent que la méthode de coloration de graphe est fiable, robuste à diverses contraintes et garantit une réponse en temps réel au tri et classement de courriers. Les taux de reconnaissance sur une base d'apprentissage de *512* documents affirment la supériorité de la coloration de graphe avec *97%* de bonne classification pour la coloration de graphe contre *89,7%* pour les *SVM* et *83,3%* pour les *k*-moyennes (figure 1.10).

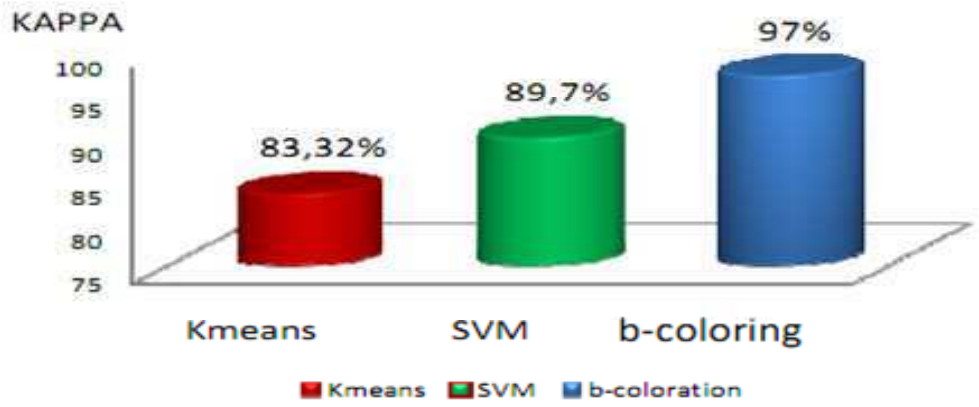


Figure 1.10. Evaluation de la classification *(Gaceb et al. [2009])*

Les trois classificateurs ont été aussi testés sur une base de test de *576* documents classés en *14* catégories qui ont été apprises et *2* nouvelles classes inconnues (de rejet) qui n'ont pas été utilisées lors de l'étape d'apprentissage. La figure 10 donne le taux de reconnaissance pour les *14* classes connues et les *2* classes inconnues. La coloration de graphe donne de meilleurs résultats en termes de taux de reconnaissance et de rejet (figure 1.11).

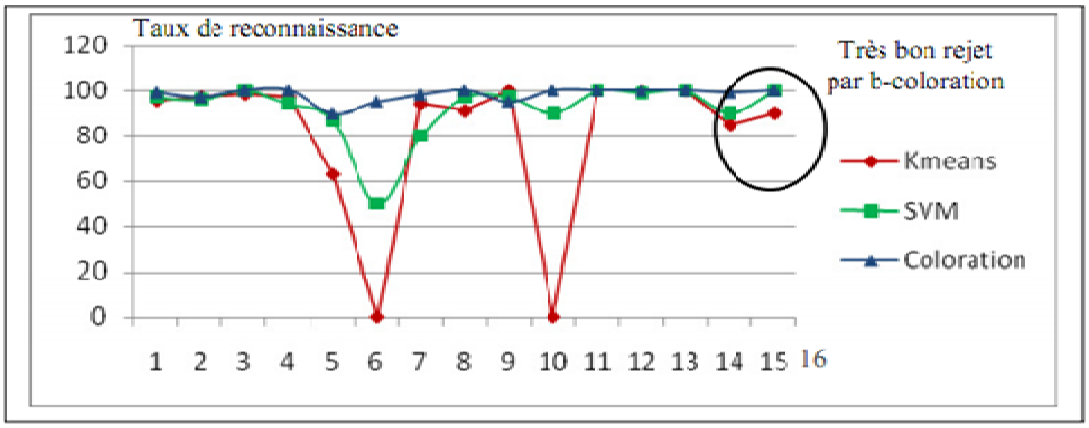


Figure 1.11. Comparaison des trois classifieurs *(Gaceb et al. [2009])*

La coloration a donné de meilleures performances de classification que les k-moyennes et que les *SVM*, mais ce n'est pas assez pour faire le choix de cette méthode de classification dans le cas de notre problématique car nous avons remarqué dans les résultats de classification de manuscrits que la plupart des classifieurs fournissent des résultats comparables en termes de taux de bonne reconnaissance.

Puisque nos travaux se positionnent sur une approche de description très locale des traits d'écriture (au niveau des fragments de traits considérés comme des primitives de l'écriture), nous remarquons que la coloration est bien adaptée à la classification de ces petits éléments

structurels *(Gaceb et al. [2009]).* Elle nous permet d'atteindre une séparation efficace des objets (fragments) sans nécessiter une connaissance a priori sur le nombre de classes et sans avoir besoin de régler les paramètres de classification. C'est pour ces raisons que nous avons décidé de choisir cette méthode de classification non-supervisée par coloration de graphe.

5 Conclusion

Nous avons montré dans ce chapitre différentes méthodes de classification de manuscrits utilisées dans le domaine de l'identification de scripteurs et la discrimination de styles d'écritures, puis nous avons illustré sur des exemples concrets emprunts de la littérature les principales familles de méthodes de classification supervisée et non-supervisée. Nous avons montré que les méthodes décrites dans ces deux familles peuvent être divisées en sous-familles. Pour chaque méthode illustrée dans l'état de l'art nous avons montré ses avantages et ses inconvénients et avons tenté des approches comparatives lorsque les conditions expérimentales le permettaient : taille et contenu des bases d'images traitées, nature des descripteurs choisis et classifieurs utilisés.

Pour achever cette représentation nous avons choisi de développer les méthodes de classification non-supervisée qui sont plus adaptées à notre problématique, et plus spécifiquement nous avons focalisé notre présentation sur l'algorithme de coloration de graphe que nous utiliserons dans le suite de nos travaux et, qui a montré des avantages très intéressants en terme d'utilisation (facilité de paramétrage) et de souplesse de classification en mettant en œuvre une optimisation de critères inter et intra classe pour la convergence du partitionnement. Pour compléter notre état de l'art sur l'analyse de styles (discrimination et identification), nous allons envisager dans le chapitre suivant les méthodes usuelles de caractérisations employées dans le domaine de la classification des manuscrits et l'identification de scripteurs.

Chapitre 2 : État de l'art sur les méthodes de caractérisation des écritures globales, locales et mixtes

Résumé: *Dans ce chapitre nous présentons un état de l'art sur les trois grandes approches de caractérisation des écritures : globales, locales et hybrides. L'objectif est de relever les différents types de caractéristiques existant dans le domaine et de justifier notre approche de la caractérisation des écritures liées à l'étude des corpus d'images du Moyen Age.*

Mots-clés: *extraction de caractéristiques, approche globale, approche locale, approche hybride.*

1 Introduction

La caractérisation des documents est une étape essentielle à toute application portant sur les éléments de contenu, parmi eux on note l'identification du style et la reconnaissance du scripteur. Selon le but à atteindre, l'analyse visera différentes informations et différentes formes de caractérisation des éléments du contenu.

Si le but est l'identification du scripteur ou la reconnaissance de style alors l'analyse va se concentrer sur l'écriture elle-même. En revanche si le but est de permettre l'accès au contenu du document, plusieurs autres informations peuvent être ajoutées comme par exemple des informations sur la mise en page ou bien sur les composantes graphiques présentes dans les images de manuscrits et dont l'analyse permettra de faciliter l'accès direct. Dans tous les cas, la caractérisation des documents est une étape nécessaire et doit rester discriminante.

Nous présentons dans ce chapitre les méthodes les plus significatives utilisées dans le domaine de la caractérisation des écritures. Elles sont ici classées selon qu'elles sont basées sur la globalité d'un document (en s'intéressant à une information de texture par exemple) ou mesurent seulement un élément local. Nous identifions également les approches mixtes utilisant les deux types d'observations, globales et locales. Nous parlerons des limites de ces méthodes pour l'identification de manuscrits, précisant ce que chacune d'elles peut apporter à la résolution de notre problématique, enfin nous terminerons par une analyse objective de ces méthodes. L'extraction des caractéristiques visuelles consiste en des transformations mathématiques du signal représentant l'image ou en des calculs à partir d'un ensemble plus ou moins important de pixels de l'image.

Nous distinguons différents types d'approches de caractérisation :

- **Les approches globales:** ces approches s'appliquent au niveau du document dans sa totalité ou bien au niveau d'un paragraphe entier. Dans tous les cas, une approche globale permet d'exprimer une impression générale, un rendu visuel global. Selon ce mécanisme de caractérisation, les descripteurs retenus peuvent être basés par exemple sur la texture, sur l'analyse multi-résolution et multi-échelle (spectrale, structurelle) et l'analyse fractale.
- **Les approches locales:** ces approches s'appliquent sur le texte sur la base d'un découpage en fragments à un niveau plus fin, localisé autour de la figure de texte, du mot, du caractère ou du graphème. Les composantes connexes, l'analyse de courbure à partir des contours sont des exemples d'approches locales pour la caractérisation.
- **Les approches mixtes:** ces approches consistent à combiner les résultats des approches locales et globales afin d'utiliser les avantages de chacune d'elles. Nous citons par exemple l'utilisation des dictionnaires de formes construits à partir de fragments de traits extraits par une analyse locale et offrant une vue globale du manuscrit à partir d'une analyse des distributions globales des fragments.

Selon le point de vue, nous nous intéressons à des niveaux de granularité et d'échelle différents, allant du bloc (ensemble de lignes), à la ligne (ensemble de mots), au mot (ensemble de caractères), ou à la lettre ou portion de lettre (ensemble de traits) pour finir au trait (graphèmes, segments). Le diagramme de la figure 2.1 montre les différents types de caractéristiques utilisées dans le domaine d'identification de manuscrit.

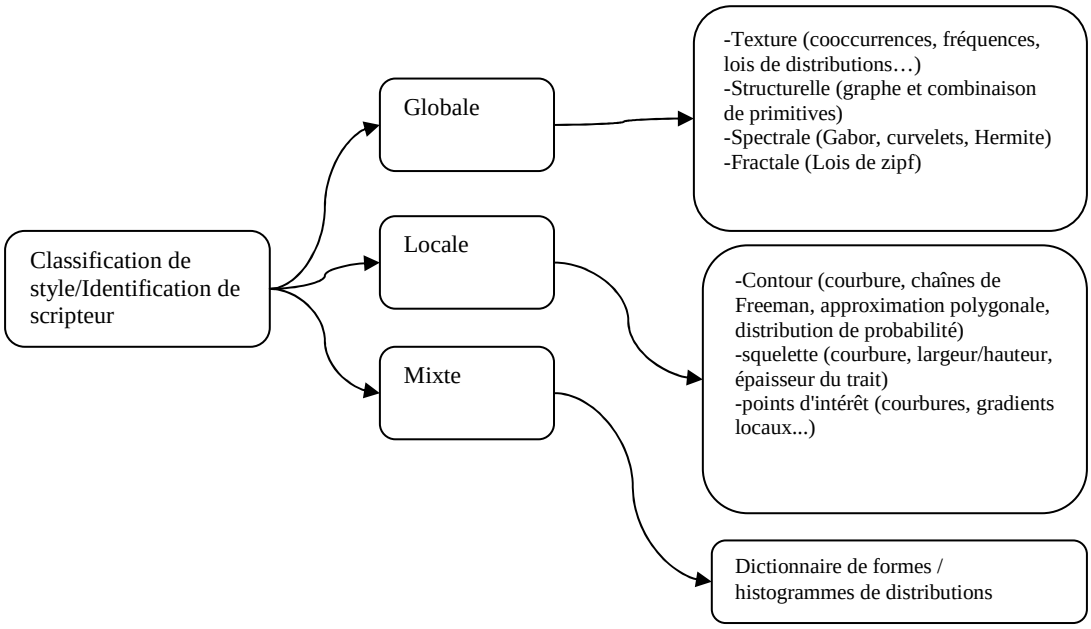


Figure 2.1. Résumé des caractéristiques extraites selon les mécanismes globaux, locaux et mixtes considérant le document à différents niveaux d'intérêt et d'échelle

Plus nous nous approchons d'une description fine, plus nous disposons d'outils de mesure et d'indicateurs calculables. Les approches de caractérisation macroscopique sont minoritaires et souvent soumises à des contraintes de taille minimale pas toujours satisfaites (il faut un minimum de lettres pour décrire une texture).

Pour caractériser un style on peut alors se ramener à considérer des points vue complémentaires.

- Prendre une signature globale exprimant les invariants de l'écriture.
- Prendre quelques éléments significatifs seulement.
- Combiner les deux sources d'informations

2 Les approches globales

Dans cette section nous allons illustrer les approches d'analyse globales : analyse de texture, analyse fréquentielle et analyse multi-résolutions et multi-échelles.

2.1 L'analyse de texture

Cette approche conduit à comparer des textes manuscrits sur la base de leur aspect global, ces méthodes sont réservées à des applications d'analyse, de classification ou de reconnaissance de style. Les caractéristiques utilisées pour la comparaison sont calculées sur l'image entière des textes soumis à l'examen. En général, ces approches s'inspirent de notions liées à la perception humaine pour la reconnaissance de styles d'écriture ou de l'écriture d'un individu (Figure 2.2).

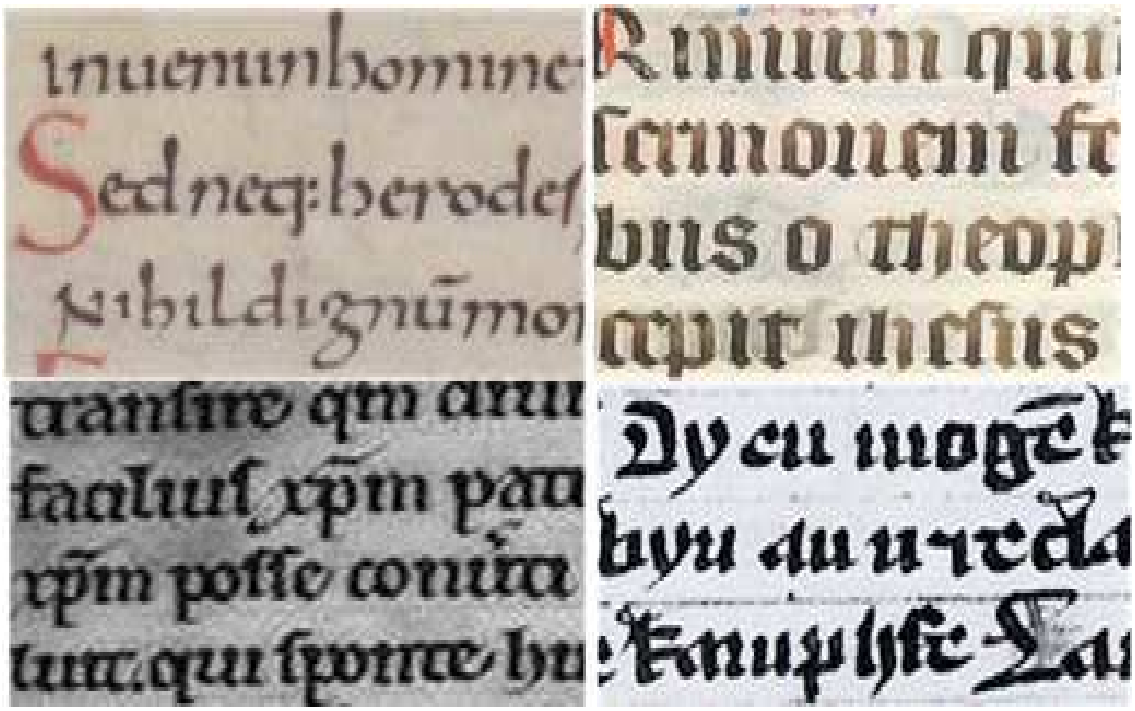


Figure 2.2. Rendu visuel de *4* styles d'écritures paléographiques (IRHT)

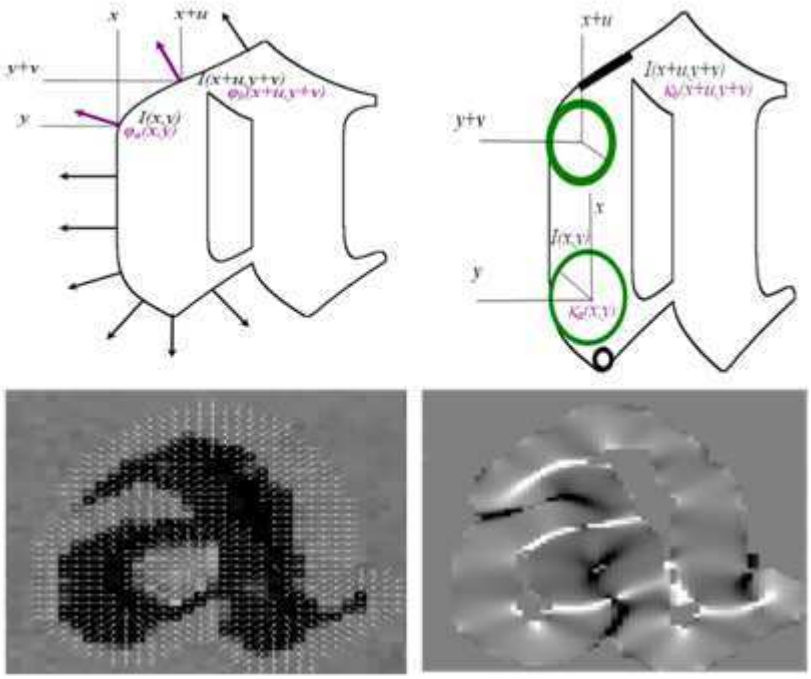
Dans *(Said et al. [2000])* les auteurs se basent sur l'analyse de texture pour identifier les scripteurs sur des écritures redressées non uniformément. Cette méthode comprend trois étapes, qui sont respectivement la normalisation, l'extraction des caractéristiques et enfin l'identification. La normalisation comprend aussi deux sous étapes, la première consiste à calculer la pente de l'écriture par une projection horizontale et à la corriger par un ajustement des lignes des composantes connexes. La seconde consiste à normaliser le texte en éliminant les espaces entre les mots et lignes pour pouvoir appliquer une analyse de texture. L'étape d'extraction de caractéristiques utilise deux méthodes distinctes comparables, un filtre de Gabor

multi-canal, et un calcul de matrice de cooccurrences. Le filtre de Gabor multi-canal extrait des caractéristiques de texture et il est représenté comme suit :

$$\begin{cases} h_e(x, y, f; \theta) = g(x, y)\cos\theta(2\pi f(x\cos\theta + y\sin\theta)) \\ h_o(x, y, f; \theta) = g(x, y)\sin\theta(2\pi f(x\cos\theta + y\sin\theta)) \end{cases} \quad \textbf{(2.1)}$$

Le *g* représente une fonction gaussienne, *f* et θ la fréquence radiale et l'orientation qui définissent la localisation du canal dans le plan fréquentiel. Dans *(Said et al. [2000])* les auteurs fixent quatre valeurs pour les fréquences (*f = 4, 8, 16* et *32* cycles) et pour chacune d'elles, 4 orientations (θ *= 0°, 45°, 90°* et *135°)* ce qui conduit à un total de *16* caractéristiques. Pour chacune des *16* caractéristiques la moyenne et l'écart type calculés forment ainsi un total de *32* caractéristiques. La deuxième méthode d'analyse de texture se base sur les matrices de cooccurrences *(Haralick et al. [1973]).* Elle présente une grande simplicité de mise en œuvre et fournit de bons résultats sur la plupart des types d'images. La matrice de cooccurrences M_C est une matrice carrée $N \times N$ où *N* représente le nombre de niveaux de gris où chaque élément de la matrice *Mc(i,j)* représente le nombre de paires de pixels séparés par une distance d et par un angle α avec l'horizontale et qui ont respectivement un niveau de gris i et j.

(Moalla et al. [2006]) ont travaillé sur *350* manuscrits médiévaux de la base de l'IRHT, dans le but d'une classification des manuscrits par style (gothique, carolingienne,…) et en se basent sur les matrices de cooccurrences. Dans leurs travaux ils utilisent quatre types de matrices de cooccurrences. Le premier type de matrices permet la comparaison des niveaux de gris de l'image avec elle-même modulo des translations et/ou rotations. Les deuxième et troisième types de matrices sont similaires au premier mais sont appliqués sur les orientations et les courbures aux bordures des traits (figure 2.3)



(a) (b)

Figure 2.3. (a) Orientation du gradient, (b) courbure gaussienne

Le quatrième type de matrice de cooccurrences est construit à partir des appariements entre l'image des orientations et l'image des courbures translatée et/ou tournée.

Quelque soit le type de matrices choisies, l'étape d'extraction des caractéristiques est la même. Un ensemble de *23* descripteurs ont ainsi été extraits des matrices de cooccurrences dont les *14* descripteurs de Haralick comme le contraste, la corrélation, la somme des carrés, l'entropie etc. et *9* descripteurs complémentaires comme la dissimilarité, la covariance, la moyenne, la probabilité maximale, etc. Ces caractéristiques on permit de donner des taux de reconnaissance des styles d'écritures variant entre *93,75%* et *100%* sur la base de l'IRHT.

2.2 L'analyse fréquentielle

Transformée de Fourier

Les premiers travaux significatifs tentant de caractériser l'écriture via une analyse fréquentielle sont ceux de *(KucKuck, [1980]).* Dans ces travaux, Kuckuck discute de quelques méthodes basées sur la transformée de Fourier (figure 2.4) ou sur la fonction d'autocorrélation, qui sont apparentées à l'analyse spectrale. Des caractéristiques sont extraites par calcul de coefficients dérivés du spectre de puissance de la transformée de Fourier.

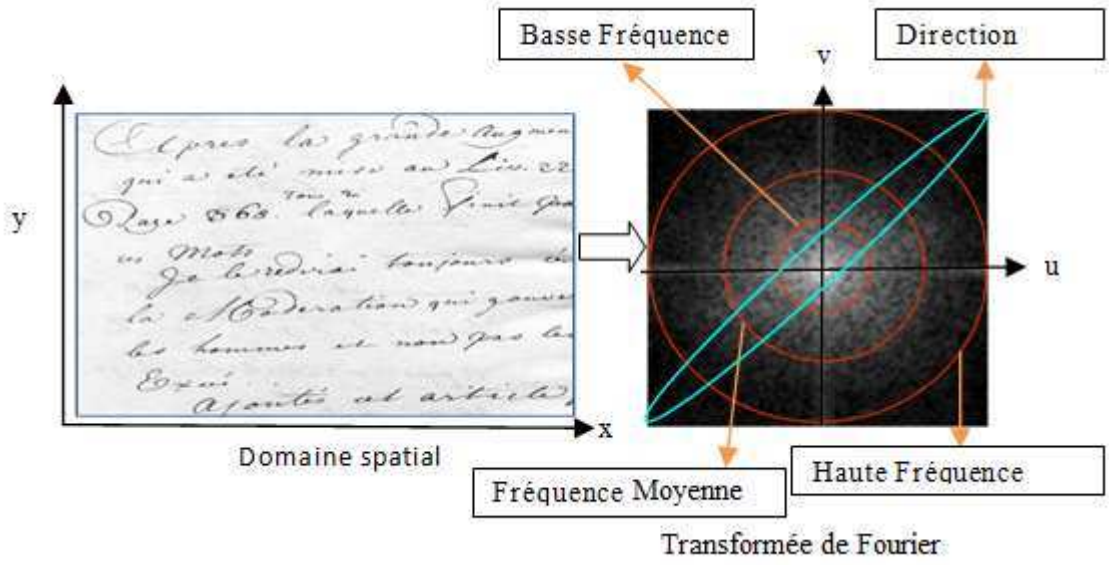


Figure 2.4. Exemple de la transformée de Fourier d'un document manuscrit

L'auteur applique également la fonction d'autocorrélation sur ce spectre pour extraire d'autres caractéristiques (figure 5). D'autres travaux, ceux *de (Dargenton, [1991])* basés sur la transformée de Fourier montrent que l'on peut accéder à d'autres informations grâce à l'analyse de la transformée de Fourier d'une portion de document (figure 2.5).

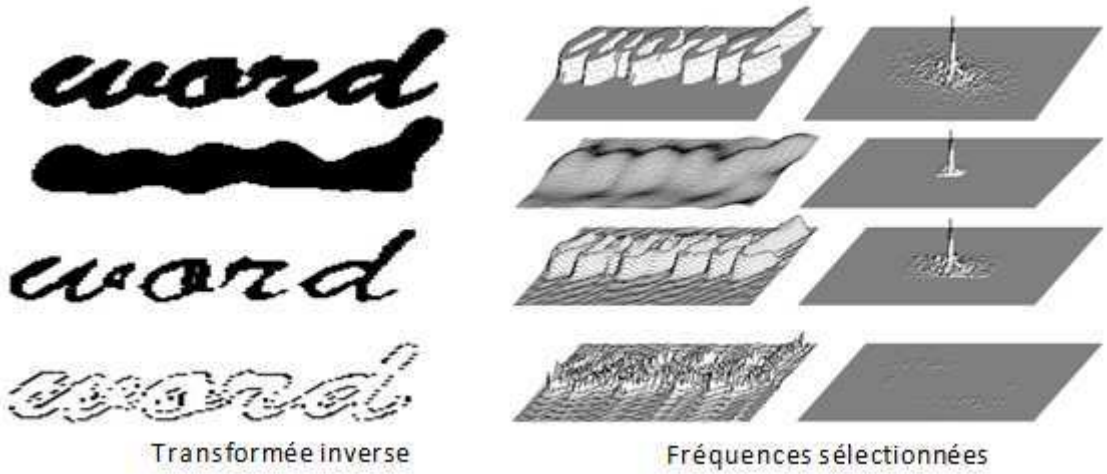


Figure 2.5. Exemples de filtrages par la transformée de Fourier et des fréquences sélectionnées *(Dargenton, [1991])*

Dans *(Yan et al. [2009])*, les auteurs ont proposé une extraction de caractéristiques spectrales basées sur la transformée de Fourier pour l'identification de scripteur. Leur méthode a été appliquée à deux bases de *100* et *500* manuscrits écrits et représentant une grande diversité d'auteurs chinois. Un ensemble de *240* caractéristiques basées sur la transformée de Fourier ont été extraites pour chacun des manuscrits des deux bases. Sur les deux bases, les taux de

reconnaissance sont respectivement *98%* et *95%*. Cette méthode donne de bons résultats à condition que les échantillons écrits contiennent un nombre suffisant de caractères chinois.

Filtres de Gabor

La fonction élémentaire de Gabor a été introduite par Dennis Gabor *(Gabor, [1946])*. Les filtres de Gabor sont utilisés dans de nombreuses applications d'analyse d'images, de détection de traits ou de contour, dans les approches de classification de texture et de compression. Un filtre de Gabor agit selon une forme gaussienne, pour cela il est considéré stable par rapport à plusieurs transformations *y* compris la translation, la rotation et la mise à l'échelle.

Un filtre de Gabor est défini par la formule suivante :

$$\psi_{s,d}(x,y) = \psi_{\vec{k}}(\vec{z}) = \frac{\left\|\vec{k}\right\|}{\delta^2}.\exp\left(-\frac{\left\|\vec{k}\right\|^2.\left\|\vec{z}\right\|^2}{2\delta^2}\right)\times\left[\exp\ i\vec{k}.\vec{z} - \exp\left(-\frac{\delta^2}{2}\right)\right] \quad \textbf{(2.2)}$$

Où $\vec{z} = [x, y]$ est une variable dans le domaine spatial et le vecteur de fréquence qui détermine l'échelle et l'orientation des filtres de Gabor.

Dans *(Eglin et al. [2006])*, les auteurs ont proposé une approche globale et sans segmentation pour la réduction de bruit et la classification de manuscrits. Leur méthode consiste à développer des outils bien adaptés pour l'amélioration de l'écriture, la séparation du fond, texte, dessins et enfin la caractérisation des formes avec des caractéristiques basées sur l'orientation. L'analyse des images est faite dans le domaine spectral par les décompositions de fréquences (Transformée de Hermite) et l'utilisation des filtres de Gabor pour l'extraction sélective d'informations. La signature est ainsi l'expression de la fonction de quantification de la densité de traits selon les différentes orientations (figure 2.6 (a)). Les maxima locaux de la courbe sont donc significatifs de la prédominance d'une orientation sur les autres. Les valeurs minimales traduisent enfin la contribution des autres directions dans la formation du trait de l'écriture.

Plus la courbe est horizontale plus l'écriture présente des courbures régulières et vice versa, des densités angulaires très contrastées présentant des pics très bien localisés peuvent provenir d'une forte inclinaison locale des traits. La figure 2.6 (b) présente *10* exemples de signatures sur un corpus des manuscrits de Montesquieu.

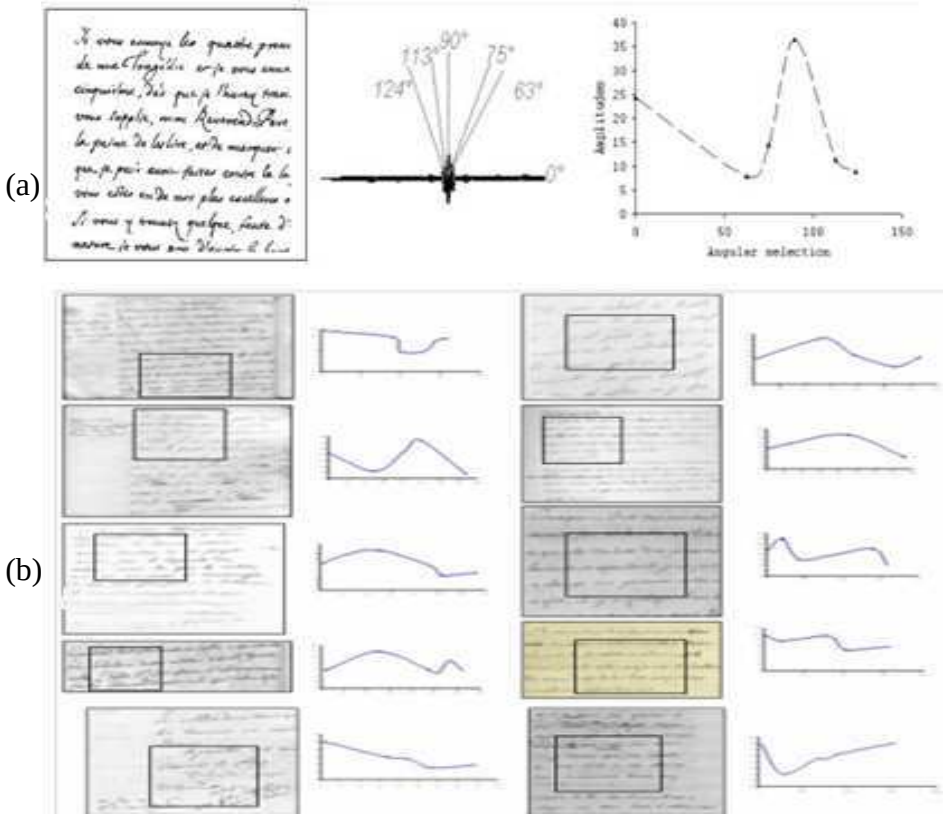


Figure 2.6. (a) Exemple de signature d'un extrait de manuscrit en 6 directions principales, (b) signature de *10* scripteurs sur des extraits du corpus de Montesquieu (qui comporte plus de 30), *(Eglin et al. [2006])*

La méthode a été testée sur des manuscrits des $18^{ème}$ et $19^{ème}$ siècles de la BNF (Bibliothèque Nationale de France). Dans leurs travaux, les auteurs se sont focalisés sur l'étude de la caractéristique de l'orientation qui révèle des propriétés globales et locales de l'écriture, avec un taux de reconnaissance de *93%.*

Dans *(Sahabi et Rahmati, [2006])*, les auteurs se sont basés sur les filtres de Gabor multi-échelle et ont utilisé les bancs de filtres de Gabor pour l'identification de scripteurs. Pour cela, un banc de filtres de Gabor selon *8* orientations et 3 fréquences λ = *(2,8;4,2;5,6)* sont utilisés. La réponse des filtres de Gabor représente des régions qui sont dans la direction et la fréquence de l'échelle. La méthode a été testée sur une base privée de *40* scripteurs et des manuscrits persans et arabes avec des taux de reconnaissance entre *56,25%* et *90%.*

Dans *(Tan [1998])*, l'auteur a proposé un ensemble de caractéristiques invariantes à la rotation basées sur la transformée de Fourier et les filtres de Gabor. Ces caractéristiques sont calculées sur un banc prédéfini de filtres de Gabor. Pour chaque fréquence, un ensemble de cinq coefficients est acquis. Six langues ont été utilisées pour montrer la pertinence du système (chinois, anglais, grec, russe, persan, et en malayalam).

Dans *(Chaudhury et. Sheth, [1999]),* les auteurs ont également exploité les filtres Gabor pour l'identification des manuscrits. Leur méthode a été testée sur *20* échantillons composés de *4* langues différentes : anglais, indien, télougou et malayalam présentant des taux de reconnaissance respectivement de *88,89%, 100%, 100%* et *90%.*

On peut constater selon le domaine d'application visé (type de manuscrits, de langues, de contenus…) que les performances des systèmes de reconnaissance de l'écrit basés sur les bancs de filtres directionnels de Gabor ne sont pas identiques. La principale limitation de ces approches est liée au fait que l'approche globale de ce type de transformées masque les particularités locales des formes, les spécificités des écritures (faible différenciation de langues).

2.3 L'analyse multi-résolution et multi-échelle

L'analyse multi-résolution a été introduite par (*Meyer, [1989]),* elle constitue un outil de traitement du signal, elle permet de décomposer un signal sur plusieurs résolutions et de le reconstruire à partir des éléments de cette décomposition.

Dans le cadre de l'identification de scripteurs, les travaux de Eglin et Volpilhac-Auger ont illustré l'intérêt d'étudier les écritures manuscrites suivant différentes échelles d'observation. Dans ces travaux, quatre dimensions sont exploitées *(Eglin et al. [2004])* :

- la direction évaluée à partir d'une analyse fréquentielle multi-échelle des images, analyse basée sur les bancs de filtres de Gabor.
- la cursivité définie à partir de l'évolution multi-résolution du nombre apparent de connexités présentes sur les lignes de texte.
- la sinuosité calculée comme une déformation de type multi-résolution des profils haut et bas des tracés.
- la complexité exprimée comme une entropie de distributions multi-échelle du tracé.

La cursivité des tracés est étudiée à partir des modifications inférées par l'application de filtrages gaussiens directionnels itératifs qui vont agir dans la direction du tracé sur la connexité des formes. Nous observons la présence de trois zones. Dans la première zone plus les lettres sont resserrées plus il y a une décroissance forte, cela signifie qu'une seule convolution gaussienne est suffisante pour fusionner un grand nombre des lettres. Nous observons aussi que la deuxième zone est plus étendue et moins pentue quand les espaces inter-mots sont plus grands et donc il faut plus de convolutions gaussiennes pour les fusionner dans un nombre réduit de blocs. La troisième zone enfin indique la fusion des espaces inter-mots. La courbe de

la figure 2.7 montre donc que le nombre de convolutions gaussiennes multi-échelle est un très bon indicateur de variabilité des distances entre les éléments textuels : caractères, mots, etc.

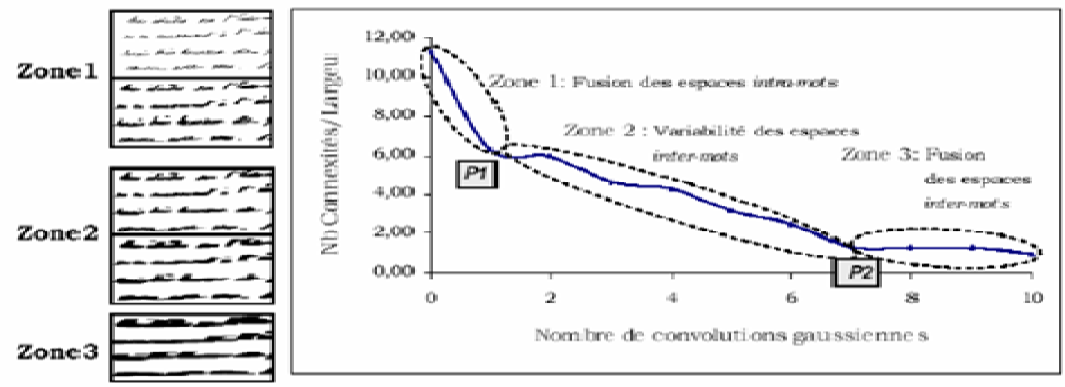


Figure 2.7. Évolution multi-résolution de la cursivité d'une portion de texte *(Eglin et al. [2004])*

L'idée principale de l'analyse multi-échelle est de mettre en évidence les différentes tailles de structures dans l'image. Pour atteindre cet objectif, différents types de filtres et de convolution ont été proposés. Les méthodes reposent généralement sur les approches itératives décomposant le signal en éléments d'information d'échelles variables. Une représentation multi-échelle consiste alors à générer, à partir d'une scène, une liste d'images à différentes résolutions spatiales. Une image à une résolution donnée s'obtient en appliquant un filtre passe-bas sur l'image pleine résolution de base. L'analyse multi-échelle utilise le concept de pyramide.

La pyramide est une représentation multi-échelle qui est construite récursivement. On distingue deux types de pyramides : les pyramides Gaussiennes et les pyramides Laplaciennes. Avec la pyramide gaussienne l'image originale subit une convolution avec un filtre gaussien : l'image résultante est une version de l'image originale sur laquelle un filtre passe-bas a été appliqué. La fréquence de coupure peut être contrôlée en utilisant le paramètre. La résultante Laplacienne est ensuite calculée comme la différence entre l'image originale et l'image après application du filtre passe-bas. Ce processus se poursuit jusqu'à obtenir un ensemble d'images filtrées selon un critère d'arrêt. En se basant sur l'esprit de la représentation multi-échelle dans *(Joutel el al. [2008])*, les auteurs proposent d'utiliser la transformée en Curvelets qui est une nouvelle représentation multi-échelle adaptée aux objets présentant de fortes structures orientées et courbes et qui constituent un outil puissant pour l'extraction des orientations et des courbures. Sur chaque sous-bande directionnelle seule l'information cherchée est marquée fortement (figure 2.8). Les segments ne correspondant pas à l'orientation considérée sont écartés. Toutefois, certains pixels de transition entre plusieurs orientations peuvent être repérés dans plusieurs sous-bandes. Dans ce cas le coefficient le plus élevé entre les différentes sous-

bandes est conservé et la sous-bande correspondante comme l'orientation dominante du trait en ce pixel est considérée

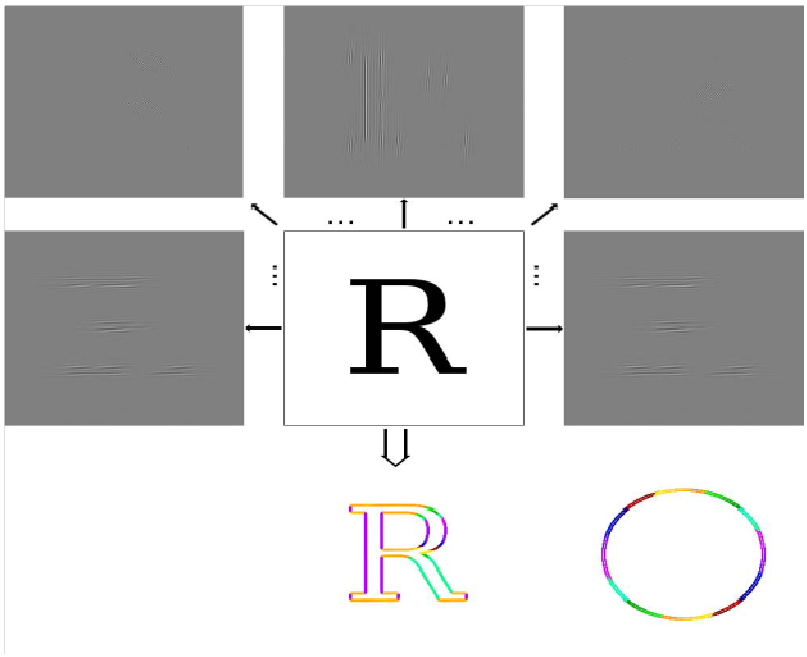
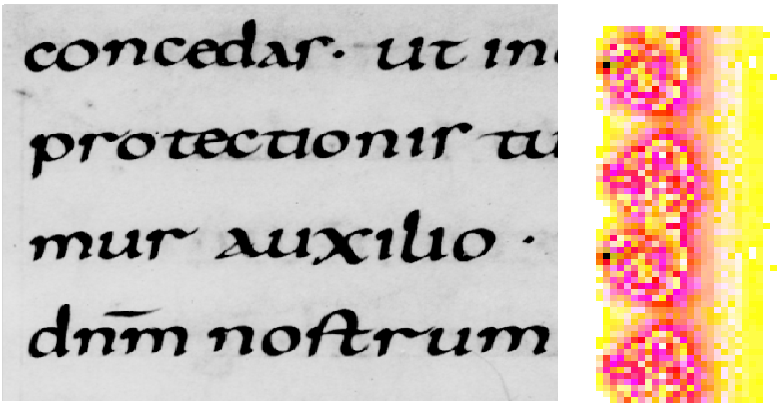


Figure 2.8. Représentation des courbures et orientations de la forme "*R*", *(Joutel et al. [2008])*

A partir des deux dimensions de direction et de courbure les auteurs proposent de construire une signature en deux dimensions exprimant leur distribution (ligne et colonne). Il s'agit d'une matrice d'occurrences des couples (courbure, orientation) représentée sous forme d'image sur la figure 2.9. Cette signature est unique pour un manuscrit et servira de caractéristique de l'écriture.



(a)Manuscrit de la base IRHT (b) signature du manuscrict

Figure 2.9. Signature de l'écriture proposée par *(Joutel et al. [2008])*

En utilisant cette technique *(Joutel et al. [2008])*, les auteurs ont produit un outil générique de caractérisation de manuscrits permettant de les classifier en familles de scripteurs. Bien que la classification et les tests de reconnaissance des scripteurs effectués sur deux bases de *400*

images : une base de textes humanistes et une base de textes du Moyen Age (composées de manuscrits de la British Library, *http://www.bl.uk* et du scriptorium numérique, *http://sunsite.berkeley.edu*) aient donné de bons résultats en précision, cette approche comporte la faiblesse de ne prendre en considération que les informations d'orientation, de courbure et de cooccurrence qui ne suffisent pas à un classement précis des textes anciens lorsque ceux-ci présentent des différences peu significatives. L'absence d'indications structurelles est un frein à leur exploitation massive pour produire un système d'identification de scripteurs robuste.

Citons aussi les travaux de *(Imdad et al. [2007])* où les caractéristiques sont basées sur la transformée de Hermite. Ils montrent que les caractéristiques basées sur la transformée de Hermite sont très utiles pour les images de texte car elles permettent d'extraire beaucoup de détails saillants selon plusieurs orientations et échelles. (*6* orientations et *4* échelles). Les expériences ont été effectuées sur la base IAM avec un taux de reconnaissance de *83%* sur un ensemble de *30* auteurs classifiés par *SVM*.

2.4. Analyse fractale

La dimension fractale, comme définie par *(Mandelbrot, [1975])*, est un nombre qui mesure le degré d'irrégularité ou de segmentation d'un ensemble et constitue de ce fait un indicateur de complexité intéressant. Le comportement fractal des écritures a été prouvé *par (Boulétreau et al. [1995])*. Des études faites plus tard montrent que sous certaines conditions, les paramètres fractals sont stables et ont un bon pouvoir de discrimination permettant de les utiliser pour une classification des écritures ou des signatures suivant le style *(Boulétreau et al. [1998]*). Le calcul de la dimension fractale est basé sur la mesure de la dimension de Minkowski-Bouligand, donnée pour un ensemble X par :

$$D(X) = \lim_{r \to 0} \left[1 - \frac{\log[A(X_r)/r]}{\log r} \right] \quad \textbf{(2.3)}$$

où *Xr* représente le dilaté de *X* de taille *r* et *A(X)* mesure l'aire de l'ensemble *X*. Pour une courbe fractale, le comportement de *log[A(Xr)/r]* en fonction de log(r) est linéaire. La limite qu'on cherche prend la valeur *1-p(X)* où *p(X)* est la pente de la courbe que les auteurs ont appelée graphe d'évolution:

$$\begin{cases} x = \log(r) \\ y = \log[\log(A(X_r)/r] \end{cases} \quad \textbf{(2.4)}$$

La figure 2.10 montre l'évolution du graphe associé à une image de manuscrit. Nous distinguons trois zones de pentes différentes, caractérisant un comportement particulier et qui correspond à une échelle d'observation particulière. La dimension fractale du manuscrit est calculée à partir de la pente de la zone *1*. Cette partie du graphe correspond aux dilatations pour lesquelles le contenu visuel du texte est acceptable. La dimension secondaire *D2* est calculée à partir de la pente de la zone *2*. Cette zone du graphe correspond à des valeurs de *r* à partir desquelles le texte est caché par des dilatations.

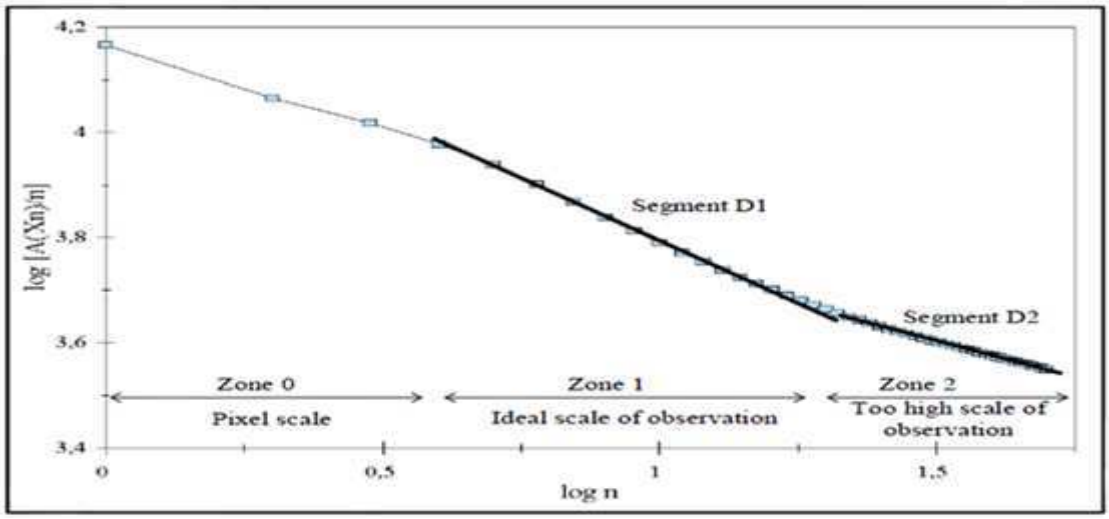


Figure 2.10. Apparence d'un graphe d'évolution *(Boulétreau et al. [1998])*

La dimension fractale reste stable en fonction de contraintes physiques liées à la fois à l'écriture et à son acquisition *(Boulétreau et al. [1998])*. Les auteurs ont vérifié que l'utilisation d'outils d'écriture divers n'a d'influence que sur les premiers points du graphe l'évolution (zone *0*), zone qui n'est pas prise en considération pour le calcul de la dimension fractale. Il a également été démontré par *(Boulétreau, [1997])* que les changements de résolution infèrent une translation des différentes zones du graphe, mais n'apportent aucune modification dans les pentes des zones *1* et *2* à partir desquelles sont calculés les paramètres. Représentant les manuscrits dans le plan, Dimension fractale vs. *D2*, les auteurs suggèrent que leur distribution est liée à la lisibilité (Figure 2.11).

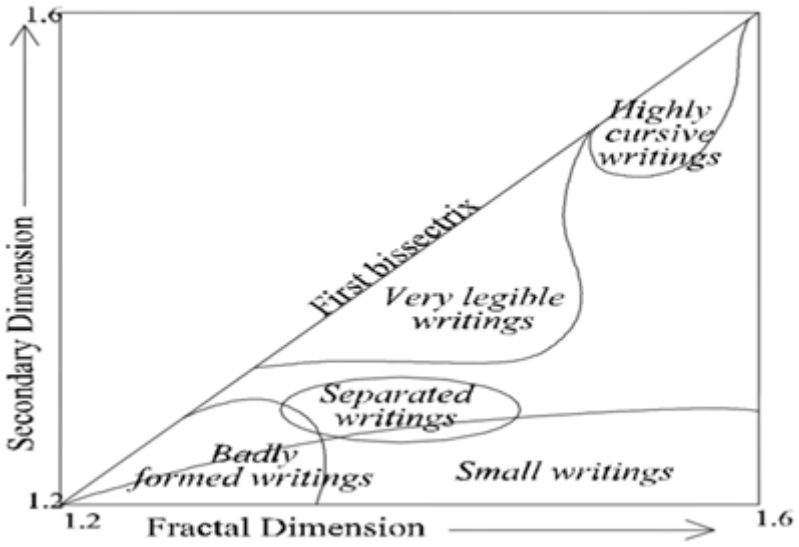


Figure 2.11. Graphe de lisibilité *(Boulétreau, [1998])*

Plus tard, des études sur l'analyse fractale de l'écriture ont été menées par *(Séropian et al. [2003])*, ainsi qu'un système d'identification de scripteurs basé sur un algorithme de compression fractale proposé par *(Fisher, [1995])*, ainsi qu'une étude proposée par *(Séropian et al. [2004])* où le but est de différencier entre des parties de texte manuscrit écrit en utilisant des alphabets différents.

2.5 Analyse par la loi de Zipf

La loi de Zipf est une loi empirique fondée sur une loi de puissance. Cette loi énonce que dans un ensemble de symboles structurés de façon topologique, la distribution des fréquences des motifs n'est pas aléatoire. La fréquence d'apparition de ces motifs M_1, M_2,…, M_n est notée par N_1, N_2,…,N_n. Quand les motifs sont triés par ordre décroissant de leur fréquence d'apparition, ils sont liés aux rangs de ces fréquences *(Pareti et Vincent, [2006])*. Cette relation est formulée par :

$$N_{\sigma_{(i)}} = k \times i^{a} \qquad \textbf{(2.5)}$$

Nσ(i) représente le nombre d'apparitions d'un motif de rang *i*, *k* et a sont des constantes. Cette loi est caractérisée par l'exposant a, k, lui, est plus lié à la longueur de la séquence de symboles étudiée. La complexité de la loi tient au fait qu'elle n'est pas linéaire. Toutefois l'application d'un opérateur logarithmique conduit à une relation linéaire caractérisée elle aussi par deux paramètres. Un codage est nécessaire quand on veut appliquer cette loi sur des images. Dans *(Pareti et Vincent, [2006])* les auteurs ont choisi de quantifier les valeurs de niveau gris

selon k niveaux, *k* étant fixé à *9*, puis à *3*. Plusieurs cas ont été étudiés, un motif *3x3* et *k = 9* ou un motif de *4* connexité et *k = 3*, produisant respectivement *99* et *35* modèles possibles. Les courbes de Zipf sont construites et approximées par des segments, justifiant la loi considérée. Les auteurs ont choisi de prendre en compte dans chaque courbe jusqu'à trois zones linéaires. La segmentation est opérée de manière récursive. Les trois pentes et abscisses des extrémités de chaque segment sont considérées comme des caractéristiques et représentent des documents. La comparaison entre deux images est faite à partir de la distance de Hamming (figure 2.12). La méthode testée sur des documents du 16ème siècle, produit un taux d'identification proche de *80%.*

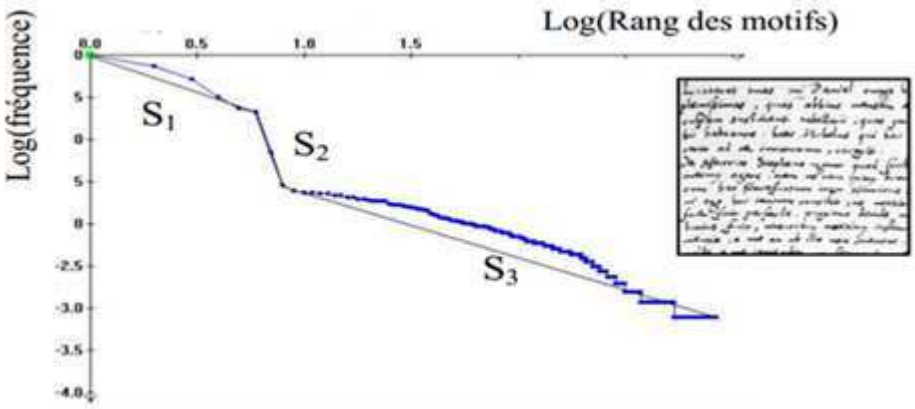


Figure 2.12. Courbe de Zipf associée à un manuscrit avec extraction des zones linéaires conduisant au calcul des paramètres de la loi, caractéristiques de l'écriture, *(Pareti et Vincent, [2006])*

Nous avons illustré dans cette section l'application des approches globales. Ces approches peuvent servir à des tâches de classification et de reconnaissance de styles ou de scripteurs mais plus efficacement de style dans la mesure où elles ont tendance à moyenner les valeurs et à ne pas tenir compte des spécificités de l'écriture.

Elles sont plutôt orientées sur des considérations visuelles générales, sur le rendu global, l'impression générale. Nous résumons aussi dans le tableau 2.1 les méthodes citées dans cette section avec les caractéristiques utilisées, la base sur laquelle les tests ont été faits, la taille de la base, l'objectif de l'étude et enfin le taux de reconnaissance. Les taux moyens de reconnaissance de style ne dépassent que très rarement les *95%,* ils sont cependant assez stables (pas de valeurs aberrantes ni d'écarts types très élevés entre les différentes méthodes proposées même si celles-ci sont appliquées à des bases de natures différentes). Cela traduit la pertinence de ce type de caractérisation mais qui pour pouvoir conduire à des résultats précis nécessitent des considérations graphométriques plus locales. La section suivante présente les familles d'approches de caractérisation basées sur une analyse fine des formes.

Tableau 2.1. Résumé des caractéristiques globales utilisées dans le domaine de l'identification de scripteurs et de la classification de manuscrits

Auteur(s)	Caractéristiques	Base	Taille de la base	Objectif	(%) de reconnaissance
(Moalla et al. [2006])	Matrices de cooccurrences	Privé-IRHT	*311*	Classification de manuscrits	*93.7%-100%*
(Yan et al. [2009])	Descripteurs de Fourrier	Privé-chinois	2 base-*100* et *500*	Identification de scripteurs	*95%-98%*
(Eglin et al. [2006])	Coefficients directionnels de Gabor	BNF	-	Identification de scripteurs	*93%*
(Sahabi et Rahmati, [2006])	Coefficients directionnels de Gabor	Privée	*40*	Identification de scripteurs	*56.25%- 90%.*
(Joutel et al. [2008])	Coefficients maximaux de Curvelets (orientation/courbure)	Humaniste-médiévale	*400*	Classification de manuscrits	*90%*
(Imdad et al. [2007])	Coefficients de la transformée polynomiale de Hermite	IAM	*300*	Identification de scripteurs	*83%*
(Seropian, [2003])	Analyse Fractale	Privée	*50* *20* classes	Identification de scripteurs	*85%*
(Pareti et Vincent, [2006])	Zipf	Textes italiens médiévaux	-	Identification de scripteurs	*80%*

3 Approches locales et description statistique

Les approches locales s'attachent à l'étude d'éléments particuliers. Ces points d'intérêts peuvent être divers selon les études. Nous allons considérer ici les principaux, les contours, leurs approximations polygonales, les points d'un squelette ou tous les points. Autour de ces points d'intérêt, différentes caractéristiques peuvent être étudiées concernant l'orientation, la courbure ou une forme locale.

3.1 Analyse par les contours

La détection de contours dans les images a commencé à partir d'opérateurs locaux, comme les opérateurs de détection de gradient ou bien les convolutions de l'image par des masques caractéristiques des contours *(Haralick et Shapiro, [1985]).* Dans les années *80*, des approches plus méthodiques ont été introduites par (*Marr et Hildreth, [1980]),* puis par Canny *(Canny, [1986]),* pour obtenir des contours plus spécifiques, fins et unitaires. Ces travaux ont donné un cadre méthodologique ainsi que la possibilité d'extraire des propriétés fines relatives aux caractéristiques du contour. Au delà des approches analytiques proposées par Canny, on peut citer également les contours actifs de *(Kass et al. [1987])* et *(McInerney et Terzopoulos, [1996]),* ainsi que les approches par ensembles de niveaux (level sets) (*Osher et Sethian, [1988]), (Chan et al. [2000]), (Vese et chan, [2002])* ou les méthodes par watershed (*Salman, [2006])*. La difficulté de la détection de contour réside dans la présence de bruit dans l'image.

Pour les approches basées sur le gradient, la première étape consiste à calculer la norme du gradient en tout point de l'image. La seconde étape consiste à extraire les maxima locaux du gradient dans la direction du gradient et la troisième est un seuillage par hystérésis appliqué à l'image pour extraire les points qui forment le contour.

Pour les approches basées sur le Laplacien, la première étape consiste à calculer le Laplacien de l'image. La deuxième étape porte sur la recherche des passages par zéro du Laplacien, la troisième étape consiste enfin en la création d'une image des passages par zéro affectés de la norme du gradient, puis dans la dernière étape un seuillage par hystérésis est appliqué sur toute l'image pour extraire le contour. Aussi notons que le filtre ou l'opérateur utilisé pour faire la dérivation a une grande influence sur les résultats.

Pour les méthodes basées sur les contours actifs, l'idée générale est de faire évoluer un contour dans l'image jusqu'à ce qu'il converge vers les contours qui nous intéressent. Le contour est contrôlé par une fonction d'énergie dont le minimum caractérise les contours d'intérêt. Nous distinguons quelques types de contours actifs comme par exemple : les modèles discrets (triangulations, modèles polygonaux), les modèles paramétriques (B-splines, superquadriques …) et les modèles de surfaces implicites (contours actifs géodésiques). Il faut noter que la dépendance à l'initialisation du contour est généralement forte, de même, plus on ajoute de contraintes plus l'optimisation devient difficile. A partir du contour, on peut extraire plusieurs types de caractéristiques, parmi ces caractéristiques citons par exemple la courbure et le codage de Freeman.

La courbure

La courbure est une caractéristique très importante de contour pour juger la similarité entre deux formes. Elle possède des caractéristiques perceptuelles importantes et s'est avérée être très utile pour la reconnaissance de formes.

Notons *K(n)* la fonction de courbure, il existe plusieurs définitions. Dans (*Mokhtarian et al. [1996])* et *(Jalba et al. [2006]) K* est défini comme :

$$K(n) = \frac{x'(n)y''(n) - y'(n)x''(n)}{\left(x'(n)^2 + y'(n)^2\right)^{3/2}} \qquad \textbf{(2.6)}$$

Par conséquent, il est possible de calculer la courbure d'une courbe planaire à partir de sa représentation paramétrique. Si *n* est le paramètre s normalisé de la longueur de l'arc, alors l'équation précédente devient :

$$K(n) = x'(s)y''(s) - y'(s)x''(s) \qquad \textbf{(2.7)}$$

Cette équation montre que la fonction de courbure est calculée uniquement à partir de dérivées analytiques, et, par conséquent, elle est invariante aux rotations et aux translations. Toutefois, la mesure de courbure dépend de l'échelle, elle est inversement proportionnelle à l'échelle. Une manière possible d'atteindre l'indépendance à l'échelle est de normaliser cette mesure par la courbure moyenne absolue.

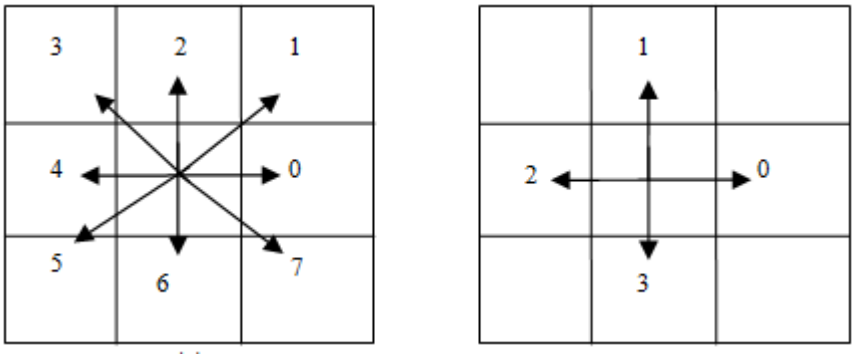
$$K'(s) = \frac{K(s)}{\frac{1}{N}\sum_{s=1}^{N}|K(s)|} \qquad \textbf{(2.8)}$$

Où *N* est le nombre de points sur le contour normalisé.

Dans *(Bulacu et al. [2007])* les auteurs ont utilisé la courbure calculée à partir du contour pour l'identification de scripteur sur des manuscrits arabes. Les tests ont été réalisés sur les *350* scripteurs avec *5* échantillons par scripteurs de la base IFN/ENIT. Le taux de reconnaissance était *99%* de bonne identification. Dans *(Tomai el al. [2004])* les auteurs ont aussi utilisé la courbure pour l'identification de scripteur. Les tests sont appliqués sur une base contenant environ *75000* images de mots écrits par *1000* scripteurs avec un taux maximum d'identification de *63%.* Les caractéristiques basées sur la courbure ont montré aussi une performance efficace dans la reconnaissance de caractères *(Legault et Suen, [1992]), (Miura et al. [1997]).*

Les chaînes de Freeman

Freeman a introduit un code de chaîne qui décrit le mouvement le long d'une courbe numérique ou bien d'une séquence de pixels en utilisant ce qu'on appelle la *8*-connexité ou la *4*-connexité. La direction de chaque mouvement est codée à partir du système de numérotation suivant : $\{i|i=0,1,2,...,7\}$ ou $\{i|i=0,1,2,3\}$ désignant un angle dans le sens anti-horaire de *45°* × *i* ou *90°* × *i* par rapport au demi-axe des *x* positifs, comme on le voit dans la figure suivante :



(a) (b)

Figure 2.13. (a) code de Freeman avec *8* directions (*8*-connexité) ; (b) code de Freeman avec *4* directions (*4*-connexité)

En codant la position relative, plutôt que la position absolue du contour, les chaînes de Freeman sont invariantes aux translations. Nous pouvons apparier deux formes en comparant les chaînes de Freeman associées, mais avec les deux principaux problèmes: 1) sensibilité au bruit; 2) invariance aux rotations. Pour résoudre ces problèmes, les codes de chaînes différentielles (DCC) et les codes de chaînes de ré-échantillonnage (RCC) ont été proposés.

Les codes des chaînes différentielles (DCC) est l'encodage des différences dans les directions successives. Cela peut être accompli en soustrayant chaque élément de la chaîne de l'élément précédent et en prenant le résultat modulo *n*, Cette différenciation permet de faire pivoter l'objet sans incréments de *90°* et de toujours comparer les objets.

Les codes de chaînes de ré-échantillonnage (RCC) consistent à ré-échantillonner le contour en une grille grossière et puis à calculer le code de Freeman à partir de cette représentation. Cela enlève les petites variations et le bruit.

Nous citons aussi les Codes de chaîne de Sommet (VCC) qui ont été proposés par *(Bribiesca, [1999])* et qui présentent d'importantes propriétés comme l'invariance à la translation, la rotation et la possibilité de représenter des formes composées de cellules triangulaires, rectangulaires et hexagonales.

Pour la méthode d'identification de scripteurs proposée par *(Siddiqi et Vincent, [2009]),* les auteurs ont extrait des caractéristiques basées sur les chaînes de Freeman. Un ensemble de quatre descripteurs (histogrammes normalisés) sont extraits à partir des manuscrits et les tests ont été appliqués sur *650* scripteurs de la base IAM et *225* de la base RIMES avec des taux d'identification de *93%* sur la base IAM et *95%* sur la base RIMES.

Dans *(Siddiqi et Vincent [2009])* les auteurs ont aussi utilisé les caractéristiques basées sur les chaînes de Freeman pour la classification de manuscrits anciens. Les tests ont été appliqués sur la base IRHT comprenant *310* manuscrits avec un taux de reconnaissance maximum de *99%.*

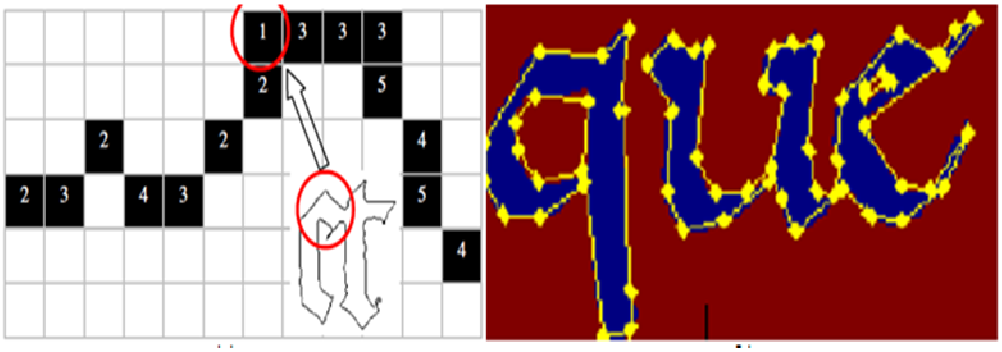
3.2 Analyse par approximation polygonale

Ces caractéristiques visent à ne garder que les caractéristiques importantes de l'écriture en rejetant les détails mineurs. Parmi ceux qui ont travaillé sur l'approximation polygonale citons *(Ramer, [1972])* qui utilise une méthode itérative pour l'approximation. Le but de l'algorithme est, étant donné une courbe composée a priori de segments de ligne, de trouver une courbe similaire avec moins de points anguleux. L'algorithme de comparaison est basé sur la distance maximale entre la courbe réelle et la courbe simplifiée. La courbe simplifiée peut être résumée par un sous-ensemble des points de la courbe d'origine.

(Wall et Danielsson, [1984]) proposent un algorithme d'approximation polygonale qui nécessite un paramètre T défini par l'utilisateur pour contrôler l'exactitude de l'approximation. Pour de grandes valeurs de T, l'algorithme crée moins de segments au prix de la dégradation de la forme du caractère et vice versa.

Cet algorithme est utilisé par *(Siddiqi, [2009])* avec $T = 2$, où ils procèdent tout d'abord par une binarisation du contour en utilisant la méthode d'Otsu, puis ils effectuent ensuite la détection des composantes connexes et pour chacune de ces composantes ils extraient les contours. Dans (*Siddiqi, [2009])* on a procédé à l'extraction des caractéristiques qui sont basées (figure 2.14) sur

- Le code de Freeman du contour.
- Un ensemble de polygones approximant le contour.



(a) (b)

Figure 2.14. Représentation du contour à partir de:(a) code de Freeman, (b) Polygone

Dans *(Siddiqi, [2009])* les auteurs ont aussi utilisé les caractéristiques basées sur l'approximation polygonale pour l'identification de scripteurs sur les bases IAM et RIMES avec des taux de reconnaissance de *87%* sur la base IAM et *88%* sur RIMES.

Parallèlement aux approches basées sur le contour, l'alternative de description des écritures porte sur une analyse du squelette des traits.

3.3 Analyse basée sur le squelette

Le squelette est très important pour la représentation d'un objet et de nombreuses approches portant sur la reconnaissance de caractères, le CBIR, dans les images de textes exploitent le squelette. La représentation basée sur un squelette n'est autre qu'une abstraction d'une forme ou bien d'un objet. Le squelette contient à la fois des caractéristiques structurelles et topologiques de l'objet original. Nous pouvons classer les points du squelette en différents types comme suit *(Lakshmi et Punithavalli, [2010]) :*

- Points simples : Un point simple est un point de l'objet qui peut être enlevé sans avoir un effet sur la topologie d'un objet.
- Points réguliers : Les points réguliers sont les points qui ont exactement deux voisins.
- Points de fin de trait : Points de fin de trait ont exactement un voisin.
- Points de jointure (ou bifurcation) : Les points de jointure sont les points qui ont exactement trois voisins.
- Points de jonction : il s'agit des points où les courbes se rencontrent et qui peuvent avoir trois (ou plus) voisins. Un point de jonction devrait satisfaire les propriétés suivantes du squelette *(Bergevin et Bubel, [2004]) :*

    a) Il doit conserver les informations topologiques de l'objet original.

    b) La position du squelette doit être précise.

    c) Il doit être stable aux petites déformations comme les saillies.

    d) Il doit contenir les centres de boules maximales, qui sont utilisées pour la reconstruction de l'objet d'origine.

    e) Il doit être invariant à la rotation et translation de l'objet.

A partir du squelette nous pouvons extraire différentes caractéristiques comme par exemple : la rondeur de l'objet (courbure vs linéarité), sa taille (largeur, hauteur), le nombre moyen de traits verticaux et horizontaux (transition pixel noir – pixel blanc sur une colonne de l'image), la courbure maximale, la position relative du précédent objet et du suivant.

Parmi les travaux portant sur l'exploitation du squelette citons les travaux *(Rousseau et al. [2004])* où les auteurs se basent sur l'analyse du squelette pour reconstruire l'ordre du tracé et parvenir à la reconnaissance de l'écriture. La méthode est constituée de trois étapes :

La première étape consiste à chercher les posers et levers du crayon partant du principe que le tracé est généralement effectué de gauche à droite et de haut en bas.

La deuxième étape consiste à reconstruire le tracé en utilisant les informations extraites à partir des posers et levers du crayon. Cet algorithme est inspiré de *(Kato et Yasuhara, [2000]),* à chaque intersection de squelette il consiste à relier les bons traits.

La troisième étape consiste enfin à trouver pour chaque couple de posers et levers quel est le meilleur tracé de l'ensemble des tracés générés en utilisant des critères absolus comme le sens de parcours des boucles et des critères relatifs comme la minimisation de la courbure globale.

Les tracés qui ne respectent pas les critères absolus sont écartés, plus spécifiquement, les tracés ne respectant pas le sens de parcours des boucles dirigées vers le haut, la droite ou la

gauche sont éliminés. La courbure des parties parcourues plusieurs fois est calculée. Le tracé est écarté si la valeur de la courbure dépasse un certain seuil.

Dans ce contexte citons les travaux de *(Pervouchine et Leedham, [2005])* qui utilisent le squelette pour l'identification de scripteur. Dans leurs travaux, un ensemble de *25* caractéristiques basées sur les squelettes de la forme (courbure, hauteur, largeur …) ont été extraites. Les tests d'identification sont appliqués sur la base CEDAR contenant *150* manuscrits écrits à partir de *30* scripteurs différents avec un taux d'identification de *98%*.

Nous avons illustré dans cette section les méthodes de caractérisation calculées localement. Ces méthodes sont plus adaptées aux tâches de classification de scripteurs. Elles permettent de montrer les spécificités de l'écriture au lieu d'avoir une vue globale du document.

Les caractéristiques locales sont des caractéristiques discriminantes calculées à partir des portions d'un trait et sont très importantes dans le domaine médico-légal d'identification de scripteur puisque les experts de ce domaine utilisent des parties distinctives des caractères comme preuves pour démontrer la ressemblance ou la différence entre deux documents *(Pervouchine et Leedham, [2007]).* C'est à partir du squelette de la lettre qu'ils extraient *25* caractéristiques structurelles pour identifier le scripteur.

Le tableau 2.2 résume les méthodes locales citées dans cette section avec les caractéristiques utilisées, la base sur laquelle les tests ont été faits, la taille de la base, l'objectif de l'étude et enfin le taux de reconnaissance.

Tableau 2.2. Résumé des caractéristiques locales utilisées dans le domaine de l'identification de scripteurs et de la classification de manuscrits

Auteur(s)	Caractéristiques	Base	Taille de la base	Objectif	(%) de reconnaissance
(Bulacu et al. [2006])	Courbure	IFN/ENIT	*350 scripteurs*	Identification de scripteurs	*99%*
(Tomai et al. [2004])	Courbure	-	*1000* scripteurs	Identification de scripteurs	*9%-63%*
(Siddiqi, [2009])	Chaîne de Freeman	IAM/RIMES	IAM : *650* scripteurs RIMES : *225* scripteurs	Identification de scripteurs	IAM : *93%* RIMES : *95%*
(Siddiqi, [2009])	Polygone	IAM/RIMES	IAM : *650* scripteurs RIMES : *225* scripteurs	Identification de scripteurs	IAM : *87%* RIMES : *88%*
(Siddiqi et Vincent, [2009])	Chaîne de Freeman	IRHT	*310* manuscrits	Classification de manuscrits	*31%-99%*
(Pareti et Vincent, [2006])	Zipf	Textes italiens médiévaux	-	Identification de scripteurs	*80%*
(Pervouchine et Leedham, [2005])	Squelette	CEDAR	150 manuscrits	Identification de scripteurs	*98%*

4 Approches mixtes et analyse statistique des décompositions locales des écritures

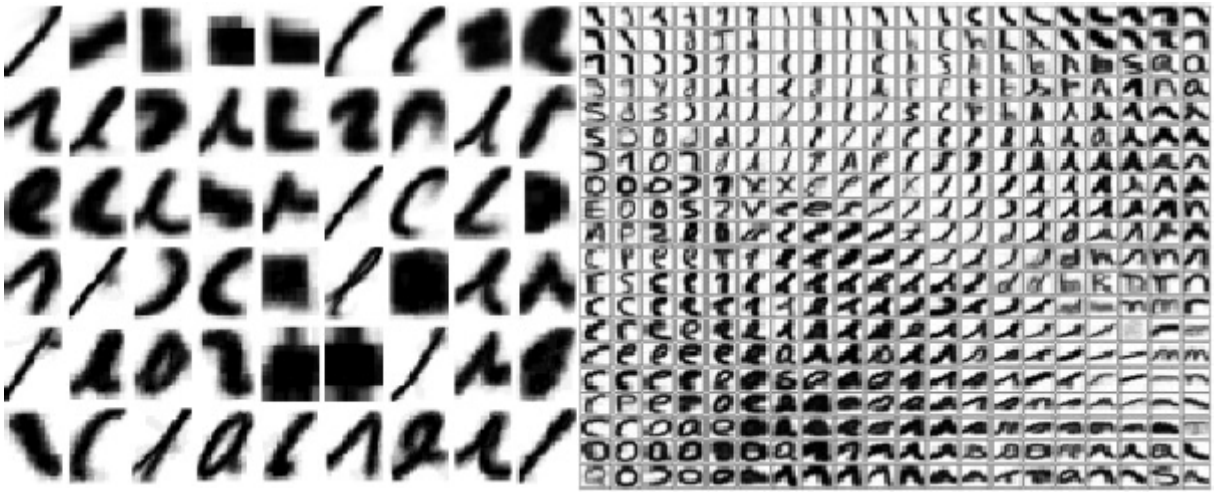
Dans cette section nous ne considérons que les méthodes basées sur un dictionnaire de formes élémentaires, même si ce ne sont pas les seules approches mixtes mais nous les avons considérées comme les plus prometteuses et celles dont nous nous sommes inspiré dans la suite de nos travaux.

Récemment, les techniques ayant pour but la reconnaissance de style et l'identification de scripteurs, se sont dirigées vers les méthodes basées sur les dictionnaires de formes où un manuscrit est segmenté en graphèmes qui sont comparés avec des éléments d'un dictionnaire soit propre au scripteur *(Bensefia et al. [2002])*, soit universel *(Bensefia et al. [2005b]), (Schomaker et Bulacu, [2004]), (Bulacu et Schomaker, [2005]).* Ces méthodes ont montré une grande performance pour l'identification de scripteurs. De même, la combinaison des caractéristiques basées sur les dictionnaires avec d'autres caractéristiques de niveaux différents a permis d'obtenir une amélioration des taux d'identification *(Bulacu et Schomaker, [2007]).*

Dans un dictionnaire, les graphèmes qui possèdent des formes similaires sont regroupés dans une même classe (Figure 2.15) et chaque graphème est décrit par un vecteur de caractéristiques comme par exemple la hauteur, la largeur, les moments de Zernike, les caractéristiques basées sur le contour, etc. Ce regroupement est fait à partir des algorithmes de clustering comme les *k*-moyennes, *kppv*, et celui des cartes de Kohonen, etc (cf. chapitre 1).

Dans *(Bulacu et Schomaker, [2005])* les auteurs ont comparé deux algorithmes, celui des *k*-moyennes et les cartes de Kohonen pour construire des dictionnaires. Les tests ont été appliqués sur *250* écrivains de la base FIREMAKER avec des lettres minuscules et *250* écrivains avec des lettres majuscules et *150* écrivains de la base imUnipen. Les résultats montrent que les dictionnaires construits à partir des cartes de Kohonen fournissent de meilleurs résultats d'identification de scripteurs que les dictionnaires construits à partir de l'algorithme *k*-moyennes.

De même dans *(Schomaker et al. [2007])* les auteurs ont utilisé ce concept de dictionnaire dans leur système d'identification de scripteurs à partir de *350* scripteurs de la base IFN/ENIT. Les dictionnaires sont construits à partir de *35000* graphèmes en utilisant les cartes de Kohonen et la taille des dictionnaires est fixée à *400 (20×20).* Dans ces travaux les résultats montrent aussi que les cartes de Kohonen permettent d'avoir de meilleurs taux d'identification de scripteur avec un taux d'identification *de 90%* pour les cartes de Kohonen et *89%* pour les *k*-moyennes. Donc nous pouvons dire que l'algorithme de construction des dictionnaires de formes, pour représenter un manuscrit, est très important et va avoir un effet sur les taux de reconnaissance.



(a) (b)

Figure 2.15. Exemples de dictionnaires de formes. (a) (Bensefia et al. [2005b]), (b) (Bulacu et Schomacker, [2005])

Une comparaison entre l'utilisation de dictionnaires propres à chaque scripteur et de dictionnaires universels a été faite *par (Siddiqi et Vincent, [2009])* pour l'identification de scripteur sur *650* scripteurs de la base IAM et *225* de la base RIMES. Les taux d'identification ont montré que les dictionnaires universels donnent de meilleurs résultats, *94%* pour les dictionnaires propres au scripteur contre *96%* pour les dictionnaires universels.

Parmi les méthodes d'identification de scripteur ou de reconnaissance de manuscrit basées sur les dictionnaires citons les travaux de *(Bensefia et al. [2005b])* qui ont développé une approche probabiliste en utilisant un dictionnaire de graphèmes sur les bases IAM et PSI. La précision du système était de *95%* sur la base IAM et *86%* sur la base PSI.

Dans *(Siddiqi et Vincent, [2008])* les auteurs ont proposé une méthode locale de caractérisation, basée sur les formes contenues dans des fenêtres de taille fixe, et cherchent à explorer la redondance des formes spécifiques au scripteur.

Leur méthode utilise des fenêtres glissantes de taille nxn pour décomposer le texte en imagettes. Pour assurer une meilleure représentativité des formes considérées, le positionnement des imagettes suit la direction du tracé. La technique consiste à trouver les extrémités des traits en utilisant un squelette et à positionner la fenêtre de découpage en fonction des points extrêmes.

Cette fenêtre est déplacée tout le long du tracé en appliquant des corrections de repositionnement là où c'est nécessaire. Par exemple si le squelette sort vers la droite ou bien la gauche, la fenêtre suivante est placée vers la droite ou la gauche respectivement (figure 2.16).

Figure 2.16. Découpage du mot headlines à l'aide de fenêtres glissantes

Dans leurs travaux, les auteurs utilisent des caractéristiques géométriques comme : la projection horizontale, la projection verticale, les profils inférieur et supérieur, l'orientation et l'excentricité. Les expérimentations ont été faites sur *100* écrivains choisis aléatoirement à partir de la base IAM et ont fourni un taux de reconnaissance de *92%.*

(Maaten et Postma, [2005]) ont utilisé une combinaison de simples caractéristiques directionnelles et d'un dictionnaire de graphèmes. La méthode a été testée sur *150* écrivains et la précision du système était de *97%.*

(Gilliam et al. [2010]) ont utilisé aussi des dictionnaires pour l'identification de scripteurs sur des manuscrits anglais des $14^{ème}$ et $15^{ème}$ siècles. La base de données complète contient environ 500 images écrites à partir de *51* scripteurs. Les dictionnaires ont été construits en se basant sur la méthode de *(Bulacu et Schomaker, [2007])* et en utilisant les *kppv* pour le regroupement des graphèmes. Le taux de reconnaissance maximum était de *78%.*

Dans *(Jain et Doermann, [2011])* les auteurs ont présenté une méthode d'identification de scripteurs en convertissant les K-adjacent segments extraits à partir du document en un vecteur de code de mots pour produire un dictionnaire qui est considéré comme un représentant du document. Pour le regroupement des segments similaires, ils utilisent comme technique de classification, la propagation affine.

Cette approche permet d'atteindre un taux de reconnaissance de *93%* sur *650* scripteurs de la base IAM. Les résultats montrent que la performance d'identification augmente à mesure que le nombre d'échantillons d'apprentissage augmente. Cette technique a été aussi testée sur des

manuscrits arabes de la base MADCAT comprenant *302* scripteurs avec un taux de reconnaissance de *90%.* Les résultats montrent que le dictionnaire est générique, indépendant des langues et des scripteurs de sorte qu'il n'a pas besoin d'être recréé en fonction de l'application.

Nous avons illustré dans cette section les approches de reconnaissance de scripteurs et de styles basées sur des dictionnaires qui, à partir de caractéristiques locales des graphèmes, permettent d'avoir une vue globale du manuscrit. Dans cette section nous avons aussi montré comment la combinaison des caractéristiques *(Siddiqi et Vincent, [2008]), (Bulacu et Schomaker, [2007])* permet d'améliorer les taux de reconnaissance.

Le tableau 2.3 résume les méthodes basées sur les approches mixtes qui sont citées dans cette section avec les caractéristiques utilisées, la base sur laquelle les tests ont été faits, la taille de la base, l'objectif de l'étude et enfin le taux de reconnaissance.

Tableau 2.3. Résumé des approches mixtes utilisées dans le domaine d'identification de scripteurs et de classification de manuscrits

Auteur(s)	Caractéristiques	Base	Taille de la base	Objectif	(%) de reconnaissance
(Bulacu et Schomaker, [2007])	Distribution oblique	IFN-/ENIT	*350* scripteurs	Identification de scripteur	*89%-90%*
Siddiqi et Vincent, [2008])	Projection, profil	IAM	*100* scripteurs	Identification de scripteur	*92%*
(Maaten et Postma, [2005])	Directionel	-	*150* scripteurs	Identification de scripteur	97%
(Gilliam et al. [2010])	Distribution oblique	-	*500* images*/ 51* scripteurs	Identification de scripteur	*47%-78%*
(Jain et Doermann, [2011])	K-adjacent segments	IAM- MADCAT	IAM : *650 scripteurs* MADCAT :*302* scripteurs	Identification de scripteur	93%

5 Conclusion

Comme nous l'avons montré, les méthodes de caractérisation globale sont plus adaptées aux applications CBIR de documents et à la classification de style d'écriture. Une vue globale du document peut être suffisante pour classer un document selon un critère donné et trouver sa ressemblance par rapport à d'autres documents.

En général, la plupart des méthodes globales sont plus efficaces pour effectuer une recherche dans un ensemble de documents où le nombre de scripteurs est limité. Leur performance commence à diminuer dès que le nombre de scripteurs devient important à cause des confusions qui ne peuvent pas être discriminées par une telle description globale. Dans ce cas là une description locale offre une analyse plus fine et peut résoudre ce problème car au lieu d'avoir un descripteur unique pour tout le document on aura un ensemble de descripteurs qui représentent

ses éléments constitutifs (mots, lettres ou graphème). Cela permet d'avoir une vue plus approfondie de la façon dont ces éléments sont formés. Cette caractérisation locale est déjà utilisée également dans une application de Word Spotting.

Les méthodes mixtes utilisant les deux types de représentations (locale et globale) offrent une meilleure précision par rapport aux deux types de caractérisations précédentes. Dans ce type d'approche, les méthodes basées sur les dictionnaires sont de plus en plus utilisées dans les applications de reconnaissance de style d'écriture et d'identification de scripteurs. Ces dictionnaires peuvent être universels à partir de tout l'ensemble des scripteurs ou bien définis pour chaque scripteur individuellement. Les méthodes basées sur des dictionnaires universels sont en général plus efficaces en termes de temps de calcul, mais, un nouveau dictionnaire de formes doit être généré si l'application est modifiée. Par contre les dictionnaires spécifiques aux scripteurs demandent un plus grand temps de calcul mais peuvent représenter un cadre générique indépendamment de l'alphabet étudié *(Siddiqi, [2009])*. Ces méthodes peuvent être utilisées pour l'identification du scripteur et les applications CBIR puisqu'elles prennent en considération les deux aspects, global et local.

Nous avons aussi constaté que la performance de l'identification de scripteur ou classification de style va dépendre de l'étape de construction du dictionnaire de formes. Nous avons vu dans les méthodes utilisées par *(Schomacker et al. [2007]), (Gilliam et al. [2010])* et *(Siddiqi et Vincent, [2009])* le nombre de classes et la taille de chaque classe a été fixée. Cela réduit la flexibilité du dictionnaire de formes à pouvoir s'adapter à de nouveaux manuscrits. Dans nos travaux nous ne connaissons pas le nombre exact de classes de graphèmes ni la taille du dictionnaire de formes qui va conduire à la signature du manuscrit. Donc si nous spécifions le nombre de classes et la taille de chaque classe a priori, peut être la performance de notre système va diminuer dans le cas où nous ne connaissons pas le nombre de classes. Pour cela nous avons introduit tout d'abord une méthode générique de classification non-supervisée basée sur la coloration de graphe (cf. chapitre 1) où seulement un seuil d'adjacence est demandé et le nombre de classes est conditionné à partir de ce seuil, donc à partir de cet algorithme nous résolvons le problème de spécification du nombre de classes. En second, nous avons décidé que le choix du seuil de coloration soit automatique et l'ensemble de caractéristiques qui caractérisent un graphème soit automatique dans le but de maximiser la performance de classification. C'est là où nous avons introduit les algorithmes génétiques (cf. chapitre 4) qui nous ont permis d'automatiser tout le processus de construction du dictionnaire de formes. En plus, l'exploitation de ces dictionnaires peut s'étendre à des domaines de recherche d'images de manuscrits (CBIR), à l'identification de scripteurs et aux méthodes de Word Spotting et Word Retrieval. En plus, les deux points de vue, local et global, sont pris en considération, donc nous pouvons profiter des avantages des deux méthodes et combler les désavantages.

Chapitre 3 : Une Approche structurelle pour la construction de dictionnaires de formes

Résumé: *La décomposition de l'écriture sur des images de manuscrit est une étape clé pour la compréhension, l'analyse automatique de ces manuscrits et la reconnaissance des styles ou des scripteurs. Elle nous permet de décrypter chaque mouvement de la plume et de définir la forme élémentaire du motif qu'elle a produit, appelé aussi graphème. Pour obtenir une bonne décomposition des traits d'écriture dans les images de manuscrits, l'extraction d'un axe médian de bonne qualité représente une démarche importante. Dans ce chapitre, nous proposons une approche de repérage de l'axe médian de l'écriture, directement appliquée aux images en niveaux de gris (libre de toute étape de binarisation). Cette approche est basée sur l'estimation de l'orientation calculée à partir de la diffusion du gradient et de l'épaisseur du trait calculée à partir de la distance de Chamfer en différents points du tracé. Elle est robuste aux dégradations des traits, au bruit de l'arrière plan et aux irrégularités d'imprégnation des encres dans le support papier des manuscrits. Les traits issus de la décomposition de l'écriture seront ensuite utilisés pour la construction de dictionnaires de formes qui vont être exploités plus tard pour l'identification des styles d'écritures.*

Mots-clés: *Analyse du manuscrit, ductus, dynamique du tracé, diffusion, rehaussement du contraste, matrice Hessienne.*

1 Introduction

En paléographie, l'étude de la formation des traits révèle des signes importants qui permettent de différencier les manuscrits d'une époque par rapport à ceux d'une autre. En effet, cette étude le fait au niveau du ductus. Dans la revue de l'école des chartes publiée en *1974 (Emmanuel, [1974]),* on peut lire une définition détaillée du ductus : « Le ductus est la composante dynamique de la morphologie ; celle-ci représente le portrait achevé du geste graphique dont le ductus traduit les étapes successives. Une morphologie qui ne tient pas compte du ductus est donc incomplète; c'est comme si on se contentait, pour représenter une sphère, de tracer un cercle sur un plan : ce n'est pas inexact, mais c'est franchement insuffisant. » Il est donc intéressant d'utiliser cette information très importante, la dynamique du tracé comme le sens de l'enchaînement des traits, l'angle d'inclinaison, etc. pour différencier les écritures, puisque les paléographes eux-mêmes certifient que l'exploitation de cette information peut être très utile. Mais l'extraction de cette information à partir des traits dans les manuscrits dégradés demande une étape de prétraitement pour mieux présenter les traits et

les séparer du fond. Les méthodes de prétraitement conventionnelles de binarisation conduisent à une perte d'information importante *(Lee et al. [1996b]), (Wang et Pavlidis, [1993]) et (Kim et Lee, [1998]).* Nous proposons ici d'utiliser une méthode de séparation des traits du fond sans passer par la binarisation pour ensuite décomposer les traits d'écritures en graphèmes. Cette décomposition est basée dans un premier temps sur l'extraction de l'axe médian de l'écriture et dans un second temps sur l'application des règles de décomposition paléographiques en utilisant les informations fournies à partir de l'axe médian.

Le chapitre va se dérouler comme suit : nous commençons par une présentation de l'état de l'art sur les méthodes de binarisation, de rehaussement de contraste et de squelettisation, étapes nécessaires permettant d'accéder aux traits d'écriture sur des images essentiellement bruitées car issues de l'acquisition de manuscrits anciens. Nous présentons ensuite notre méthode d'analyse structurelle basée sur l'analyse de la dynamique du tracé, pour enfin terminer par la décomposition du texte en graphèmes en suivant des règles de décomposition spécifiques aux écritures du Moyen Age et enseignées par les experts paléographes.

2 État de l'art

L'état de l'art ce compose de *2* parties : une partie de prétraitement des images de textes anciens dégradés et l'extraction de l'axe médian. La partie de prétraitement comporte deux parties abordant respectivement la binarisation et le rehaussement du contraste. Le rehaussement de contraste est une étape facilitant le repérage des bords des traits souvent confondus avec l'arrière plan, sur des supports papier ayant absorbé les encres vieillies.

2.1 Prétraitement

Les prétraitements généralement appliqués aux images de documents consistent à éliminer les défauts liés à l'image afin de faciliter l'étape de reconnaissance. Ces défauts peuvent être de deux types : ceux qui sont liés à la chaîne de numérisation (inclinaison du document, luminosité, bruit, …) et ceux qui sont liés à la qualité intrinsèque du document (les taches d'humidité, l'apparition du verso, la présence de trous, de pliures ...). Pour la correction de l'inclinaison des lignes, il existe de nombreuses approches, celle *de (Trincklin, [1984])*, celle de *(Baird, [1987])*, la transformée de *(Hinds et al. [1990])* etc. Pour corriger le problème de luminosité, Belaid dans *(Belaid et al. [1992])* utilise des méthodes basées sur la modification d'histogramme. Pour éliminer les points parasites *(Belaid et al. [1992])* utilisent un filtrage passe bas, d'autres utilisent un filtre passe haut comme Brink dans *(Brink et al. [2012])* ou un filtrage morphologique comme Cannon dans *(Cannon et al. [1999])* ou encore les décompositions d'ondelettes comme on le retrouve dans les travaux de Boulehmi dans *(Boulehmi et al. [2008]).*

Pour la séparation recto/verso, il existe deux approches principales : l'approche « aveugle » qui réalise la séparation afin de ne retenir que le recto *(Drira et al. [2006]),* et l'approche « non-aveugle » qui réalise la séparation afin d'exploiter les informations figurant sur les deux faces *(Tan et al. [2002]), (Lins et al. [1994]).*

Ces étapes de prétraitement sont importantes sur un grand nombre de documents numérisés du patrimoine, ceux-ci ayant subi de nombreuses dégradations dues au temps, les usages intensifs et les modalités de stockage. En fonction de la nature des documents sélectionnés dans notre étude dédiée au projet Graphem, nous avons choisi de ne pas les mettre en œuvre systématiquement. Les aspects de restauration ne seront pas directement exploités, seul le rehaussement de contraste local dans les zones extrémales des traits sera utilisé dans la quasi-totalité des bases étudiées. Compte tenu du fait que notre contribution porte sur la décomposition de l'écriture en traits simples exprimant un mouvement continu de plume, nous n'avons pas cherché à isoler les lignes les unes des autres, nous avons donc écarté les questions liées aux enchevêtrements de lignes et à la présence d'écritures multi orientées. Seule la contrainte directement liée au vieillissement des encres a été prise en compte et nous a donc conduit à élaborer des solutions permettant un repérage robuste à l'effacement des traits (notamment aux extrémités des tracés). C'est dans ce contexte que nous allons étudier les méthodes de binarisation et celles de rehaussement de contraste permettant une meilleure étude du trait.

2.1.1 Les méthodes de binarisation

La séparation avant/arrière plan est réalisée bien souvent par une simple binarisation sur les images de textes. Il s'agit de passer d'une image en niveaux de gris ou en couleur à une image bitonale (noir et blanc). On distingue essentiellement trois catégories de méthodes de binarisation selon la nature du seuillage utilisé : les méthodes globales, les méthodes locales et les méthodes hybrides qui exploitent les deux approches précédentes.

Les méthodes globales de binarisation

Les méthodes globales ont pour but de déterminer un seuil unique pour tous les pixels de l'image et partent du point de vue que les objets doivent avoir une distribution de niveaux de gris relativement distincte de la partie fond. Dans ce cas, la recherche de seuil s'effectue par l'analyse de l'histogramme des niveaux de gris et par la détermination d'un minimum local, les pixels ayant un niveau de gris inférieur au seuil sont mis en noir et les autres en blanc. Ces méthodes sont classifiées de la manière suivante :

- Les méthodes basées sur la séparation de distribution *(Cocquerez et Philipp, [1995]), (Fisher, [1958]).*

- Les méthodes basées sur l'analyse discriminante *(Otsu, [1979]), (Yao-Hong, [2007]).*
- Les méthodes basées sur la notion d'entropie *(Esquef et Albuquerque, [2002]), (Kapur et al. [1985]).*
- Les méthodes basées sur la transformation d'histogramme *(Sahoo et al. [1988]).*
- Les méthodes basées sur la matrice de cooccurrencess *(Kohler, [1981]).*
- Les méthodes basées sur les réseaux de neurones *(Khashman et Sekeroglu, [2008]), (Babaguchi, [1990]).*

Ces méthodes de binarisation globales ont été appliquées sur l'image originale de la figure 1, produisant des rendus visuels de qualité très variable, voir figures 3.1 et 3.2.

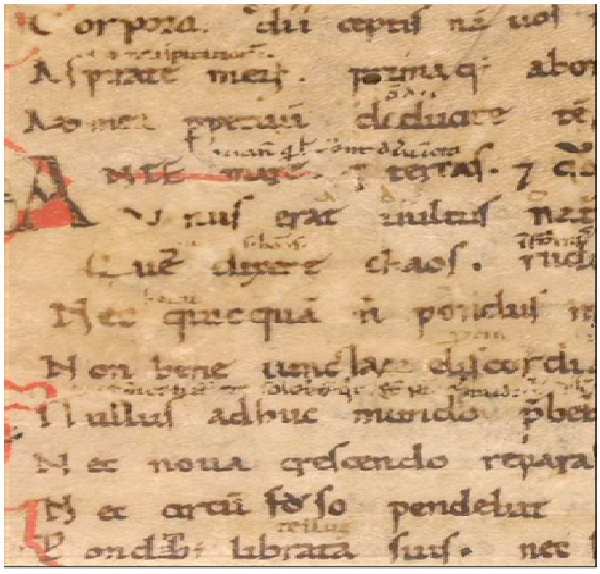
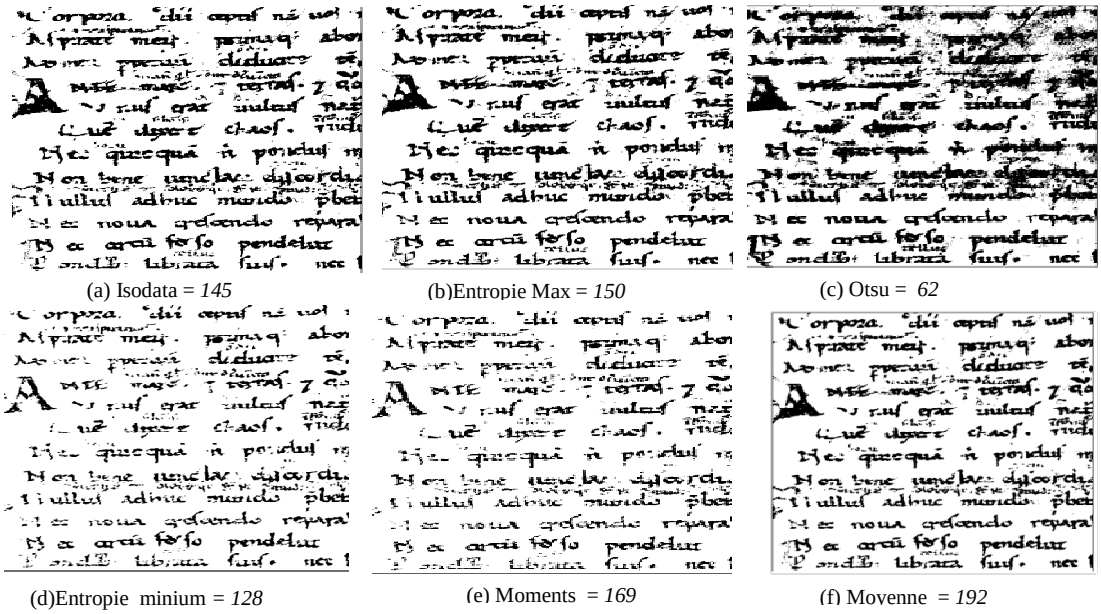


Figure 3.1. Image Originale extraite de la base de manuscrits médiévaux de l'IRHT

Au vu des résultats visuels de binarisation, nous remarquons que ce type d'approches globales n'est pas robuste aux dégradations existant dans les manuscrits médiévaux, ni aux variations de couleurs entre le premier plan et le fond. Ce phénomène s'observe particulièrement sur l'image rendue par la binarisation selon la méthode d'Otsu où nous n'obtenons pas une bonne séparation entre l'avant et l'arrière plan.



(a) Isodata = *145* (b)Entropie Max = *150* (c) Otsu = *62*

(d)Entropie minium = *128* (e) Moments = *169* (f) Moyenne = *192*

Figure 3.2. Résultats de binarisation avec un seuil global de binarisation: (a) Isodata *(Velasco, [1980]),* (b) Entropie Max *(Cheng et al. [1998]),* (c) Otsu *(Otsu, [1979]),* (d) Entropie minimum *(Li et Lee, [1993])*, (e) Moments *(Tsai, [1985]),* (f) Moyenne *(Glasbey, [1993])*

Les méthodes de binarisation locales

Les méthodes de binarisation locales (dites adaptatives) s'adaptent au contexte de chaque pixel par le calcul d'un seuil pour chaque pixel de l'image en fonction de l'information contenue dans son voisinage. Cela permet de compenser les variations de luminosité et les dégradations locales d'une image. Si la fenêtre couvre une zone de l'image faiblement contrastée, la sensibilité du seuil de détection est automatiquement augmentée. Cette adaptation aux changements locaux de contraste explique la popularité de ces méthodes sur les images de documents dégradés ou ceux qui utilisent des couleurs d'encre multiples. Habituellement, l'adaptation est obtenue en balayant l'image en zigzag par une fenêtre d'analyse centrée sur chaque pixel dans laquelle on réalise le calcul d'un seuil local. La complexité est de l'ordre de $N \times M \times P$ où $(N \times M)$ représente le nombre de pixels de l'image et P le nombre de pixels de la zone d'analyse locale.

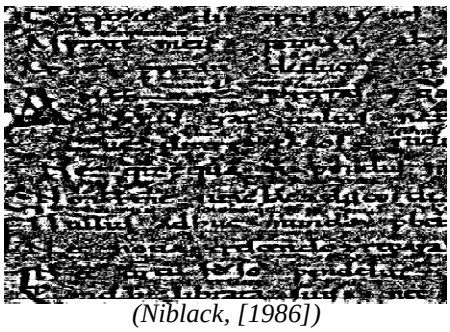
Ces méthodes sont classifiées de la manière suivante :

- Les méthodes basées sur les réseaux de neurones : (*Chigusa, [1992]), (Hamza et al. [2005]).*
- Les méthodes de seuillage local basées sur le concept de moyenne et d'écart type locaux comme dans l'approche Niblack présentée dans *(Niblack, [1986]).* Cette méthode est basée sur le calcul d'une valeur de seuillage en faisant glisser une fenêtre sur l'image. Pour chaque pixel un seuil T est calculé sur la base de certaines statistiques telles que la

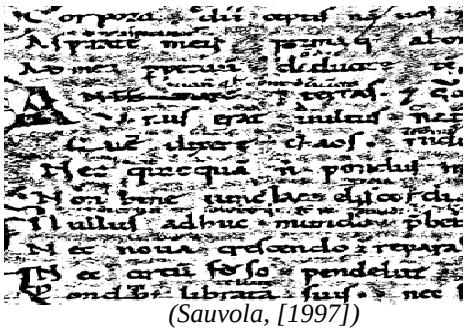
moyenne *m* et l'écart-type *s*, qui sont calculés sur les niveaux de gris des pixels voisins dans la fenêtre par la formule suivante: $T = m + k \times s$ avec *k* est une constante négative. Dans une étude comparative menée par *(Trier et Jain, [1995])*, les auteurs ont montré que la méthode de Niblack segmente bien les caractères de texte et donne ainsi de meilleures performances sur des images de documents comparativement à d'autres méthodes de binarisation globales et locales. Cette efficacité a été également confirmée par He dans *(He et al. [2005])* qui a comparé la méthode Niblack à d'autres méthodes plus récentes. Cependant, cet algorithme produit un bruit sur les images dont le fond est dégradé, ce qui entraîne la nécessité d'un post-traitement qui consomme beaucoup de temps.

- les méthodes basées sur l'étude locale des valeurs de gradients, proposées par Wolf dans, *(Wolf et al. [2002])* et *(Wolf et doermann, [2002]).*

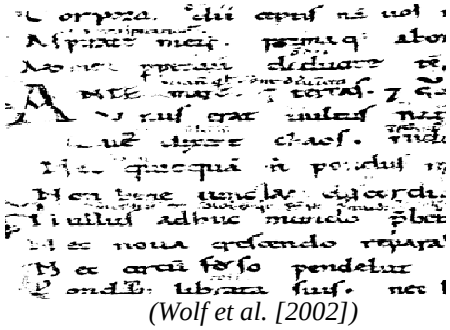
Pour les méthodes de binarisation locales, nous avons testé les méthodes de : Niblack, de Sauvola, et de Wolf (figure 3.3). L'application de ces méthodes, nous a conduit à définir des fenêtres de taille *50×50* permettant de couvrir au moins *1* à *2* caractères ainsi qu'une constante *k = 0,2* reprise de *(Wolf et al. [2002])* et qui semble fournir de bons résultats.



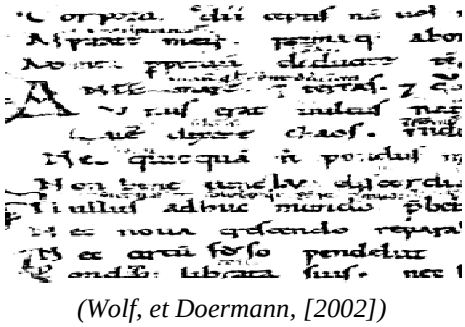
(Niblack, [1986])



(Sauvola, [1997])



(Wolf et al. [2002])



(Wolf, et Doermann, [2002])

Figure 3.3. Méthodes de binarisation locales appliquées au document de la figure 1

Dans une étude faite par *(Khurshid et al. [2009])*, les méthodes de Sauvola *(Sauvola, [1997])*, Niblack *(Niblack, [1986])*, et Wolf *(Wolf et al. [2002])* ont été testées sur des documents anciens, ils ont employé la méthode NICK qui donne aussi de bons résultats. Le système *ABBYY* Fine Reader *9* (intégrant un système de reconnaissance de caractères et de structures) a été utilisé pour tester la performance de chaque méthode. Le meilleur taux de reconnaissance a été obtenu à partir de la méthode de Wolf en comparaison avec celles de Niblack et Sauvola. Mais dans les deux cas de binarisation locale et globale, on a toujours une perte d'information malgré la meilleure performance des méthodes locales (figure 3). Cette perte d'information est préjudiciable au bon suivi du tracé en construisant une mauvaise décomposition du tracé en graphèmes. C'est pour cela que nous ne pouvons pas nous appuyer sur ces méthodes dans l'étape de prétraitement.

Les méthodes de binarisation hybrides

Ces méthodes utilisent en même temps les deux types de binarisation globale et locale. (*Trier et Taxt, [1995])* ont proposé une méthode adaptée aux documents techniques comme les cartes. Pour cela, ils calculent le Laplacien des dérivées partielles issues des résultats du filtre médian sur l'image initiale, créent une image intermédiaire labélisée puis procèdent à la séparation entre le fond et la forme par l'analyse des labels des pixels. En dernière étape, ils appliquent un post-traitement présenté dans *(Yanowitz et Bruickstein, [1989])* pour améliorer la qualité de l'image binaire. Cette méthode produit de bons résultats sur des documents de bonne qualité, mais présente l'inconvénient de devoir fixer plusieurs paramètres qui sont difficiles à régler en pratique.

Dans *(Kavallieratou et Stamatatos, [2006])* les auteurs ont proposé une méthode de binarization hybride dans le but d'améliorer la qualité des manuscrits. Au début, une méthode de binarisation globale est appliquée sur toute l'image. Ensuite, les zones de l'image qui contiennent encore des bruits de fond sont détectées et la même technique est de nouveau appliquée à chaque zone séparément. Par conséquent, l'algorithme assure une meilleure adaptation dans le cas où différents types de bruits coexistent dans différentes zones de la même image, tout en évitant le coût de calcul et le temps d'application d'une binarisation locale sur toute l'image. Les résultats sur des documents historiques ont montré que cette méthode est robuste au bruit et que la qualité des documents dégradés a été améliorée tandis que les documents en bon état n'ont pas été affectés. Dans *(Chang, [1995])* on a proposé d'égaliser l'histogramme du fond pour faciliter la distinction entre les caractères et le bruit. Puis ils ont utilisé le Laplacien pour reconstruire la forme des caractères binaires.

Dans *(Savakis, [1998])*, l'auteur propose deux méthodes de seuillage d'images adaptées à la numérisation à grande vitesse des documents. Le premier algorithme utilise un seuillage adaptatif : il applique soit un seuillage local sur les pixels qui possèdent un gradient local élevé, présentant une forte transition entre deux zones, soit un seuillage global sur les pixels homogènes appartenant au fond et possèdant un gradient faible. Le second algorithme est basé sur le suivi du tracé utilisant un regroupement basé sur une variante des algorithmes *k*-moyennes. Les deux approches peuvent être utilisées indépendamment ou peuvent être combinées pour offrir de meilleurs résultats. Dans *(Tanaka, [2009])*, l'auteur propose un post-traitement basé sur l'analyse du fond des images de documents pour corriger le résultat du seuillage adaptatif proposé par la méthode de Niblack.

Evaluation des méthodes de binarisation

Les méthodes de binarisation globale ont l'avantage d'être très rapides mais le changement d'éclairage, la présence de bruit et la dégradation de l'encre sur les manuscrits médiévaux vont réduire la qualité de la binarisation.

Les méthodes de binarisation locales dépassent ces limites et sont mieux adaptées aux changements de contraste local. En revanche, elles demandent plus de calcul, ce qui les rend plus lentes que les méthodes de seuillage global. Par ailleurs, elles peuvent conduire à des résultats de sur-segmentation des défauts et des textures du fond de l'image, et sur les images de documents textuels cela peut provoquer des difficultés à traiter les caractères dont les tailles peuvent varier puisque la taille de la fenêtre d'analyse est fixée dès le départ. Le tableau 3.1 résume les principales caractéristiques de rendu liées à l'usage de méthodes de seuillage globales, locales et hybrides.

Tableau 3.1. Résumé des méthodes de binarisation : globales, locales et hybrides

Type	Auteurs	Principe	Caractéristiques
Binarisation locale	*(Bernsen, [1986])*	Estime la valeur du seuil en faisant la moyenne de la plus haute et la plus basse valeur de la fenêtre	Le seuil est trop bas lorsque la fenêtre est centrée sur du fond
	(Niblack, [1986])	Amélioration de *(Bernsen, [1986]) :* prise en compte de la variance et de la moyenne	Même problème que *(Bernsen, [1986]):* apparition de bruit sur les zones uniformes
	(Sauvola, [1997])	Insère des constantes dans la méthode *de (Bernsen, [1986])* afin d'améliorer la méthode sur les zones uniformes	Les constantes à ajuster empêchent la méthode de traiter parfaitement des documents non uniformes.
	(Khurshid et al. [2009])	Inspirée de Niblack, le seuil de binarisation est trouvé pour chaque pixel en prenant en compte ses pixels voisins dans une fenêtre glissante	Améliore considérablement la binarisation des images de page « blanches » et claires, et dans le cas où l'image présente de faible contraste, en déplaçant vers le bas, le seuil de binarisation
	(Wolf et al. [2002])	Utilise les champs de Markov pour savoir où se trouvent les caractères	L'utilisation de *(Sauvola et Pietikäinen, [2000])* rend la technique victime des mêmes limitations que pour Sauvola.
	(Gatos et al. [2006])	Cherche à estimer le fond pour ensuite faire un seuillage sur la différence entre le fond et l'image d'origine	Bonne performance
Binarisation globale	*(Otsu, [1979])*	D'après l'histogramme, cherche à maximiser la variance intra-classe du «texte» et du «fond»	Problèmes pour les documents mal éclairés.
	(Wu et Manmatha, [1998])	Rend l'image floue pour mieux séparer l'histogramme et utilise un seuillage global	Problèmes lorsqu'il n'y a pas deux modes distincts sur l'histogramme.
	(Tsai, [1985])	Découpe l'image récursivement en quad tree et applique le seuillage d'Otsu sur chaque zone	Problème de temps de calcul et risque de générer des zones totalement noires
Binarisation hybride	*(Trier et Tax, [1995])*	«Ternarise» l'image en fonction du gradient puis utilise une heuristique pour réduire à deux classes	Bons résultats sur les images de bonne qualité, problèmes avec le réglage des paramètres
	(Wu et Amin, [2003])	Applique un seuillage global puis adapte le seuil sur les caractéristiques spatiales des composantes formées dans la $1^{ère}$ étape	Bons résultats sur des images d'enveloppes simples, bien contrastées et de bonne qualité

2.1.2 Les méthodes de rehaussement de contraste

Le rehaussement de contraste est une étape très importante sur les images dont les encres ou les tracés présentent des affadissements en bordure. Tel est le cas dans les images acquises dans des conditions très contraintes, comme les images médicales, ou les images présentant des dégradations importantes, comme on le constate sur les images de textes anciens. Récemment, on a pu observer un développement important des techniques de rehaussement de contraste dans les domaines de l'imagerie médicale, le multimédia, la transmission d'images, etc. Il est souvent considéré comme l'un des problèmes les plus importants dans le traitement d'images. *(Pei et al. [2004]), (Kaur et al. [2011])*. L'objectif du rehaussement de contraste est d'augmenter la visibilité des détails qui peuvent être altérés par l'intensité lumineuse globale ou locale de l'image. Plusieurs techniques de rehaussement de contraste existent dans le domaine. Elles peuvent être groupées en méthodes linéaires ou non linéaires *(Al-amri et al. [2010])* ou en méthodes agissant dans le domaine spatial ou fréquentiel *(Vishwakarma, [2012])*, ou encore en méthodes globales ou locales.

Nous proposons de présenter les méthodes de rehaussement de contraste comme suit :

- Les méthodes locales.
- Les méthodes globales.

Avant de commencer le classement des méthodes définissons tout d'abord les termes que nous utilisons dans cette section :

- **Méthodes du domaine spatial:** Dans ces méthodes nous travaillons directement sur les pixels. Les valeurs des pixels sont manipulées pour obtenir le rehaussement désiré.
- **Méthodes du domaine fréquentiel:** Dans ces méthodes l'image est tour d'abord transférée dans le domaine fréquentiel. Cela signifie, que la transformée de Fourier de l'image est calculée en premier. Toutes les opérations de rehaussement sont appliquées sur l'image transformée de Fourier de l'image. Une transformée de Fourier inverse est appliquée à l'image pour obtenir les résultats dans le domaine spatial.
- **Méthodes linéaires:** Ce type de rehaussement dilate linéairement les valeurs originales de l'image en une nouvelle distribution. En dilatant les valeurs de l'image, la plage totale de sensibilité du dispositif d'affichage peut être utilisée. Le rehaussement de contraste linéaire fait en sorte que les petites variations dans les données soient plus évidentes.
- **Méthodes non linéaires:** Celles-ci reposent sur des transformations non linéaires.

Passons maintenant au classement de ces méthodes suivant le critère global ou local. Nous citerons les méthodes les plus utilisées dans ce domaine, nous sommes conscients qu'il existe

un grand nombre de méthodes de réhaussement de contraste, mais ce thème n'entre pas dans le cadre des objectifs que nous visons.

2.1.2.1 Méthodes locales

Les méthodes locales de rehaussement de contraste tentent d'améliorer la visibilité des détails locaux de l'image.

Égalisation locale d'histogramme

Dans les méthodes d'égalisation d'histogramme locales qui sont des méthodes non linéaires spatiales proposées par *(Gonzalez et Wood, [2002]), (Sherrir et Johnson, [1998]), (Pizer et al. [1984])* un sous bloc rectangulaire de l'image d'entrée est d'abord défini puis un histogramme de cette région est obtenu, ensuite sa fonction d'égalisation de l'histogramme est déterminée. Le pixel central de cette région est égalisé en utilisant la fonction d'égalisation de l'histogramme. Le centre de la région rectangulaire est ensuite déplacé vers le pixel adjacent et l'égalisation d'histogramme est répétée. Cette procédure est répétée pixel par pixel pour tous les pixels d'entrée. Cette méthode permet à chaque pixel de s'adapter à sa région d'appartenance, de sorte que le contraste élevé peut être obtenu pour tous les emplacements dans l'image. Cependant, comme l'égalisation locale d'histogramme doit être effectuée pour tous les pixels dans le cadre de l'image entière, la complexité de calcul est très élevée *(Kim et al. [2001]).*

Étirement de contraste local

Dans les méthodes locales d'étirement de contraste, qui sont des méthodes linéaires spatiales, chaque plage de couleur dans l'image (RVB) est considérée. La plage de chaque couleur sera utilisée dans le processus d'étirement pour présenter chaque plage de couleur. Cela fournira à chaque palette de couleurs un ensemble de valeurs min et max *(Mokhtar et al. [2009]).*

Approche multi-échelle pour la correction du contraste *(Frangi et al. [1998])*

Cette approche de rehaussement de contraste utilisée dans l'extraction des vaisseaux sanguins peut être considérée comme une approche linéaire, elle utilise le concept de la théorie multi-échelle linéaire proposé dans *(Florack, et al. [1992])* et *(Koenderink, [1984]).* Elle repose sur une décomposition multi-échelle de la matrice Hessienne (matrice des dérivées secondes de l'image) pour la détection des vaisseaux sanguins dans les images de rétine. Les structures locales de l'image peuvent être décomposées par extraction des directions principales en utilisant les valeurs propres de la matrice Hessienne. Cette méthode utilise simultanément les valeurs propres et vecteurs propres de la matrice Hessienne pour dériver une fonction discriminante ayant une réponse maximale sur les structures de forme tubulaire qui représentent des structures linéaires. La norme de la matrice Hessienne permet de faire la distinction entre le premier et l'arrière plan. La méthode est basée sur l'observation que l'intensité des dérivées est

moins élevée pour les pixels de l'arrière plan. L'avantage de cette méthode basée sur la matrice Hessienne est qu'elle peut capturer les vaisseaux sanguins de différents diamètres en raison de l'analyse multi-échelle. La méthode de Frangi est expliquée en détails dans la section 3.2.

2.1.2.2 Méthodes globales

Les méthodes de rehaussement de contraste globales résolvent les problèmes qui se manifestent globalement comme les conditions excessives ou pauvres de luminance que nous trouvons dans l'image originale.

Égalisation d'histogramme

L'égalisation globale de l'histogramme qui est une méthode non linéaire spatiale utilise l'information de l'histogramme de l'image d'entrée entière pour sa fonction de transformation. Bien que cette approche soit adaptée pour l'amélioration de l'ensemble de l'image elle ne parvient pas à préserver les caractéristiques de luminosité locales de l'image d'entrée *(Shanmugavadivu et Balasubramanian, [2010]).*

Correction de la luminosité

Le filtre fonctionne dans l'espace couleur RVB et ajuste la luminosité des pixels en augmentant les valeurs RVB de chaque pixel par la valeur de réglage spécifié.

Remappage de couleurs

Pour chaque pixel de l'image, le filtre change ses valeurs en des valeurs qui sont stockées dans le tableau de remappage en utilisant les valeurs des pixels comme indices. Par exemple si la valeur est (*32, 96, 128*), le filtre change cette valeur en *(RedMap[32], GreenMap[96], BlueMap[128]).*

Filtres Homomorphiques

Les Filtres Homomorphiques sont considérés comme des méthodes non linéaires de rehaussement de contraste et appartiennent au domaine fréquentiel. Le filtrage Homomorphique fréquentiel agit sur les basses et hautes fréquences pour faire un rehaussement de contraste. Il utilise la caractéristique qu'une image $f(x,y)$ peut être représentée par le produit de son éclairement lumineux $i(x,y)$ qui représente les très basses fréquences et de sa réflectance $r(x,y)$ qui représente les très hautes fréquences. Le filtrage Homomorphique vise à les séparer. Tout d'abord le logarithme naturel est appliqué sur l'image avant d'extraire le spectre fréquentiel. De cette façon on sépare la composante de l'éclairement lumineux de celle de la réflectance. Chaque composante est ensuite filtrée par $H(u,v)$, un filtre atténuant les basses fréquences et rehaussant les hautes fréquences. Les transformées inverses ramènent ensuite les composantes

de l'image au domaine spatial. A la fin on applique la fonction exponentielle pour récupérer l'image filtrée et rehaussée *g(x,y) (Al-amri et al. [2010]).*

Étirement du contraste global

Les méthodes globales d'étirement de contraste qui sont des méthodes linéaires spatiales considèrent à la fois toute la plage de palette de couleur de l'image pour déterminer le maximum et le minium pour toutes les couleurs RVB de l'image. La combinaison des couleurs RVB va produire seulement une valeur pour le maximum et le minimum pour les couleurs RVB *(Mokhtar et al. [2009]).*

Dans le tableau 3.2, nous résumons les méthodes de rehaussement de contraste en indiquant à quelle famille elles appartiennent avec leurs avantages et inconvénients.

Tableau 3.2. Résumé des méthodes de rehaussement de contraste avec : G(Global), L(Local), LN (Linéaire), NL (Non linéaire), S(Spatial), F(Fréquentiel)

Méthode	G/L	LN/LN	S/F	Avantage(s)	Inconvénient(s)
Égalisation locale d'histogramme	*G*	*NL*	*S*	Automatique, Reproductible, Localement adaptative Produit généralement des images de bonne qualité.	Renforce aussi le bruit
Égalisation globale d'histogramme	*G*	*NL*	*S*	Temps de calcul faible	Traite toutes les régions de l'image avec la même importance Préservation des détails
Étirement de contraste local	*L*	*LN*	*S*	Bonne représentation visuelle de la scène originale	Perte de détail due à la saturation, la coupure et la mauvaise visibilité dans certaines régions de l'image
Étirement de contraste global	*L*	*L*	*S*		
Filtres Homomorphiques	*L*	*NL*	*F*	Peut être utilisé dans le domaine spatial	Convolution coûteuse en temps de calcul
Frangi	*L*	*L*	*F*	Robuste aux dégradations	Pas de bon traitement des bifurcations Réduction du diamètre d'une structure
Correction de la luminosité	*G*	*L*	*L*	Temps de calcul	Traite toutes les régions de l'image avec la même importance
Remappage de couleurs	*G*	*NL*	*L*	Temps de calul	Traite toutes les régions de l'image avec la même importance

2.1.2.3 Évaluation de la pertinence des méthodes de rehaussement de contraste

Pour tester la pertinence des méthodes de rehaussement de contraste, nous avons utilisé la base de DIBCO11 *(Pratikakis et al. [2011])* présentée dans ICDAR 2011 (compétition de binarisation) comme vérité terrain. L'objectif est de voir quelle méthode de rehaussement de contraste permet la meilleure amélioration des résultats de binarisation quelque soit l'approche utilisée. Pour la binarisation, nous avons utilisé la méthode globale d'Otsu. La figure 3.4 présente les résultats de binarisation selon la méthode d'Ostu obtenus sur les images ayant été traitées par plusieurs méthodes de rehaussement de contraste et les résultats sans aucun rehaussement sur les *8* images de la base de DIBCO11. La méthode de Frangi améliore les résultats de la binarisation sur les images *1, 2, 3* et *4*. En moyenne sur les *8* images le taux de bonne binarisation après rehaussement par la méthode de Frangi est *84,6%* contre *76,65%* pour la méthode basée sur la luminosité et contre *77,37%* sans rehaussement de contraste. Le taux de « bonne binarisation » est estimé à partir d'une vérité terrain fournie par la compétition de binarisation ICDAR 2011.

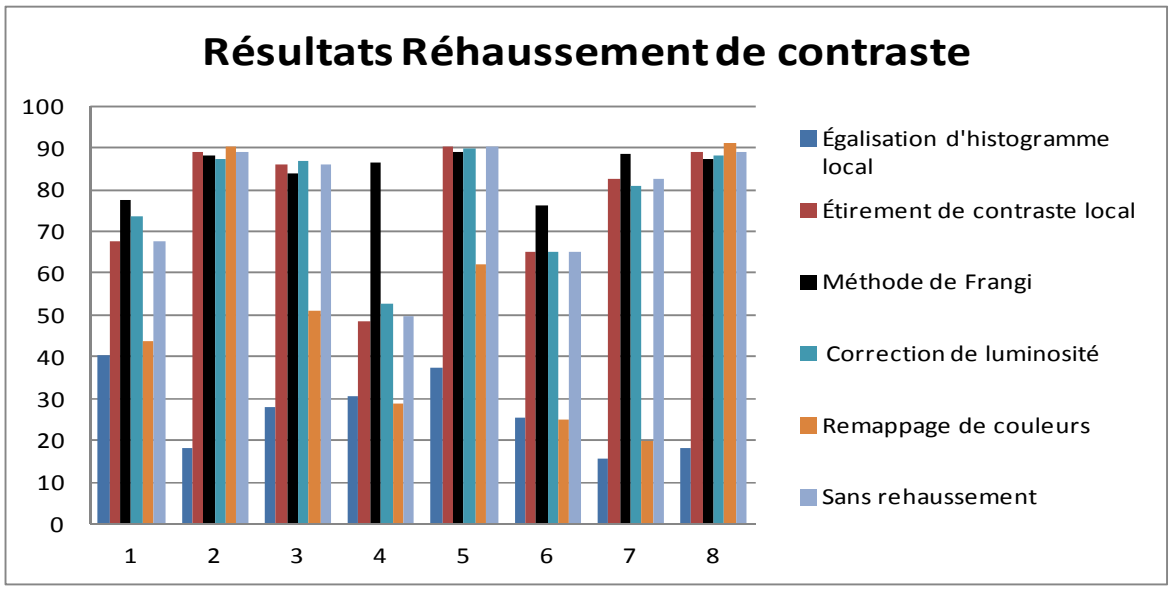


Figure 3.4. Résultats de binarisation par la méthode d'Otsu à la suite de l'application de différentes techniques de rehaussement de contraste. L'axe des abscisses présente les *8* images de la base

Nous remarquons que la méthode de Frangi offre une bonne robustesse aux dégradations et offre une bonne séparation de l'arrière et de l'avant plan par rapport aux autres méthodes de rehaussement de contraste (figure 3.5). Les raisons de tels résultats sont liées à la capacité de la méthode à rehausser les petites formes géométriques anisotropes bruitées, cette méthode ayant déjà fait ses preuves dans le domaine de l'imagerie médicale avec de très bons résultats dans le rehaussement de vaisseaux sanguins. Pour ces raisons, nous avons décidé d'utiliser la méthode de rehaussement de contraste de Frangi qui offre de bonnes perspectives dans le domaine d'analyse des documents.

Figure 3.5. Résultats des méthodes de rehaussement sur l'image 7 de la base DIBCO11

2.2 Méthodes de squelettisation

Un bon algorithme de squelettisation pour les applications de notre domaine doit posséder les propriétés suivantes *(Fan et al. [1997])* :

- Préserver la connectivité du squelette.
- Faire converger le squelette vers une largeur unitaire.
- Éviter les érosions excessives.
- Être insensible au bruit des contours.
- Avoir des temps de calcul raisonnables.

La notion de squelettisation fut introduite la première fois dans le domaine de la reconnaissance des formes par *(Blum, [1964]).* La réduction d'un objet à ses éléments essentiels peut permettre l'élimination de contours dénaturés tout en gardant des propriétés topologiques et géométriques de la forme. Centré dans la forme du tracé, le squelette fournit une représentation équivalente au tracé, mais unidimensionnelle. Le découpage de l'écriture en graphèmes peut se faire entièrement (et plus simplement) à travers celui du squelette qui contient l'essentiel de l'information synthétisée. L'extraction du squelette sur des documents anciens dégradés devient plus difficile dans la mesure où les bruits du bord et les irrégularités de contour sont des paramètres nuisibles. Il existe actuellement une grande variété de méthodes de squelettisation et le choix dépend de la nature des images. Nous groupons ces méthodes en six catégories selon la technique utilisée :

L'amincissement morphologique: l'amincissement consiste à retirer, au fur et à mesure, les points du contour de la forme, tout en préservant ses caractéristiques topologiques. Ces méthodes nécessitent une étape préalable de binarisation des images en niveaux de gris. Celle-ci conduit à une grande perte d'information lorsque les documents sont anciens et de mauvaise qualité *(Lee et al. [1995])* et *(Maio et Maltoni [1997]),* elle donne ainsi des traits binaires dégradés : caractères cassés, fusionnés ou biaisés (trous, bruit).

Le Diagramme de Voronoï: le squelette d'une forme continue est inclus dans le diagramme de Voronoï des points de sa frontière *(Schmidt, [1989]).* Cette approche est définie dans un espace continu et produit un squelette connecté. L'inconvénient de ce type de méthodes vient de la difficulté d'échantillonner les contours ce qui définit la qualité du diagramme de Voronoi et la méthode nécessite l'élagage des branches à l'aide d'étapes complexes de post-traitement *(Attali, [1995]).*

La transformée en distance: la carte de distances d'un objet consiste à associer à chacun de ses points, sa distance au point de contour le plus proche. Les maxima locaux de la carte de distance correspondent exactement aux points du squelette de l'objet. Plusieurs distances ont été utilisées dans ce cadre : Euclidienne *(Danielsson, [1980])* et *(Choi et al. [2003])*, Chamfer *(Rosenfeld, [1968])*, etc. La méthode est appliquée le plus souvent sur des images binaires mais aussi sur des images en niveaux de gris. Avec ces méthodes, l'extraction de squelette est fortement sensible aux déformations des contours de tracé qu'on rencontre souvent sur les images que nous voulons traiter.

Les heuristiques: ces méthodes reposant sur des heuristiques s'appliquent directement sur les images en niveaux de gris. Les heuristiques règlent des paramètres qui gèrent la détection de l'axe médian. Elles sont très utilisées pour l'extraction du squelette des traits sur des empreintes digitales et leurs résultats sont nettement plus robustes et efficaces sur des images dégradées que ceux des 3 familles de méthodes précédentes *(Maio et Maltoni, [1997]) (Qi, [2004])* dans lesquelles des critères complexes sont utilisés pour arrêter le processus de squelettisation et préserver le squelette. Nous retrouvons ces approches dans le domaine de l'imagerie médicale comme dans les travaux de *(Sun, [1989])* qui présente un algorithme géométrique de suivi du tracé sur les images de rayon-X. Cet algorithme est connu sous le nom de « Algorithme de Sun ». *(Aylward et bullitt, [2002])* ont adapté ce principe aux objets tubulaires (vaisseaux sanguins) sur des images *3D* en niveaux de gris. *(Yim et al. [2000])* utilisent aussi les heuristiques pour l'extraction des vaisseaux sanguins dans les images médicales en niveaux de gris. L'inconvénient principal de ces méthodes est l'absence d'auto adaptation aux changements d'orientation, de diamètre et aux problèmes de chevauchement et de croisement de vaisseaux sanguins ou de traits sur des images dégradées.

La détection des contours: ces méthodes utilisent les contours pour naviguer le long du trait et détecter l'axe médian par corrélation entre une ligne et ses deux bords. Dans ce cadre, une approche itérative intéressante est proposée dans *(Zhang et al. [2007])* pour détecter l'axe médian dans des images de neurones. D'autres méthodes, fondées sur le même principe, sont utilisées dans le suivi des routes sur des images satellitaires (Peteri et al. [2003]) et *(Poz et al. [2006])*. Ce type de méthodes ne peut pas être facilement appliqué sur des manuscrits dégradés où les contours de traits sont souvent déformés et discontinus. Dans ce cas, le suivi de contours risque de se perdre dans de petites chaînes parasites.

Les nouvelles techniques: de nouvelles techniques ont vu le jour notamment avec l'utilisation des ondelettes *(You et al. [2007]),* des level sets *(Kimmel et al. [1995]), (Mayer et al. [1998])* qui ont utilisé pour l'extraction automatique des routes, des équations aux dérivées partielles (EDP) utilisant des champs de diffusion *(Grigorishin et al. [1998]), (Chung et Sapiro [2000]),* (*Siddiqi et al. [2002]), (Yu et Bajaj, [2004]), (Pervouchine et al. [2005])* et *(Cheriet et Doré [2006]).* Les deux dernières approches permettent de squelettiser des images en niveaux de gris en évitant toutes les limites liées à la squelettisation des images binaires. Ce type d'approches nécessite plusieurs itérations de lissage, mais il est plus robuste aux dégradations. Le tableau suivant résume les avantages et inconvénients des méthodes de squelettisation citées ci-dessus (tableau 3.3).

Tableau 3.3. Avantages et inconvénients des 6 méthodes de squelettisation

Méthodes	Avantages	Désavantages
Amincissement morphologique	Conservation de la topologie et de la connectivité	Binarisation Temps de calcul
Diagramme de Voronoï	Précision Préservation de la connectivité.	Difficulté d'échantillonner les contours Complexité et temps de calcul Sensibilité au bruit
Transformée en distance	Temps de calcul	Préservation de la connectivité non garantie
Méthodes Heuristiques	Pas de binarisation Robustes aux dégradations	Le nombre important de paramètres
Détection des contours	Pas de binarisation	Sensible aux contours dégradés
Nouvelles techniques	Pas de binarisation Robuste aux dégradations	Temps de calcul Nombre de paramètres

Le tableau suivant fait le bilan des méthodes de squelettisation selon la technique utilisée, le domaine d'application et le type d'images (tableau 3.4).

Tableau 3.4. Bilan sur les méthodes de squelettisation

Auteur(s)	Méthode de squelettisation	Domaine d'application	Type d'images
(Sun et al. [1989])	Heuristiques	Imagerie médicale	Niveaux de gris
(Bullitt et al. [2000])	Heuristiques	Imagerie medicale	Niveaux de gris
(Chung et sapiro [2000])	Diffusion de vecteur, anisotropique	Lettres et images médicales	Niveaux de gris
(Kimmel et al. [1995])	Level sets	Les routes	Niveaux de gris
(Attali, [1995])	Diagramme de Vornoi	Objets tubulaires arbitraires	Binaire
(Aylward et al.[2002])	Heuristiques	Objets tubulaires	Niveaux de gris
(Cheriet et Doré ,[2006])	EDP	Reconnaissance de formes	Niveaux de gris
(Danielsson, [1980])	Transformée en distance	Reconnaissance de formes	Binaire
(Rosenfeld et Pfalz [1968])	Transformée en distance	Reconnaissance de formes	Binaire
(Choi et al. [2003])	Distance euclidienne	Reconnaissance de formes	Binaire
(Kang et Kim, [2004])	Heuristiques	Reconnaissance des chiffres	Niveaux de gris
(Ahuja et chuang [1997])	Champs de diffusion	Reconnaissance de formes	Binaire
(Grigorishin et al. [1998])	Champs électrostatiques	Reconnaissance de formes	Binaire
(Su et al. [2009])	Amincissement et graphe	Analyse de manuscrits	Binaire
(Pervouchine et al. [2005])	B-splines	Analyse de manuscrits	Niveaux de gris

3 Suivi du tracé et analyse de l'axe médian

La détermination de l'axe médian constitue une des étapes fondamentales pour l'extraction efficace des formes manuscrites. Nous présentons dans cette section l'approche de détection de l'axe médian que nous avons conçue. La méthode comporte une étape de prétraitement basée sur la méthode de Frangi, une étape de calcul du rayon à partir de la distance de Chamfer pondérée *(Meyer et Maragos, [1999])* sur l'image en niveaux de gris et enfin l'extraction de l'axe médian à partir de la diffusion du gradient au centre des zones de tracées puis une dernière étape correspondant au suivi du tracé inspiré de la méthode de Xu *(Xu et al. [2007])*, méthode très utilisée dans le domaine de l'imagerie médicale pour le suivi des vaisseaux sanguins. Les

sections suivantes détaillent les différentes étapes de cette approche d'extraction de l'axe médian et de suivi de tracé.

3.1 Principe général de notre approche de suivi du tracé

Le principe général de notre méthode de décomposition de l'écriture manuscrite en traits est présenté dans la figure 3.6. De manière à ne pas perdre d'information, notre algorithme s'applique directement sur les images en niveaux de gris sans passer par une étape de binarisation. Nous procéderons par 3 étapes successives : une étape de pré-traitement, une étape d'extraction de l'axe médian et enfin une étape de suivi du trait.

Dans **l'étape *1*** nous appliquons l'approche de rehaussement du contraste selon la méthode proposée par Frangi dans *(Frangi et al. [1998])* directement sur l'image en niveaux de gris I pour mieux mettre en évidence le tracé par rapport au fond. Puis la nouvelle image subit un lissage gaussien pour enlever les discontinuités, trous, bruits et déformations. L'image résultante est une image en niveau de gris que nous noterons I'.

Dans **l'étape *2***, après avoir appliqué le rehaussement du contraste, nous calculons la distance de Chamfer pondérée sur I' afin d'obtenir la valeur du rayon R automatiquement à chaque point du tracé *(Meyer et Maragos, [1999])*. Cette approche nous libère de tous les inconvénients liés à l'utilisation d'un rayon avec une taille fixe comme cela est le cas dans la méthode de Xu proposée dans *(Xu et al. [2007])*. La méthode du calcul de rayon basée sur la distance de Chamfer offre une très bonne adaptation au changement d'épaisseur de traits. Sur la même image I' résultant de l'étape de rehaussement de contraste, on procède ensuite à l'estimation de l'axe médian en utilisant la diffusion du gradient *(Lebourgeois et Emptoz, [2007])*. La diffusion du gradient va nous permettre d'avoir une très bonne représentation du tracé essentiellement au niveau des extrémités des formes. Comme pour le calcul du rayon selon la distance de Chamfer, nous travaillons sur l'image I' en niveaux de gris. Nous évitons ainsi de passer par une binarisation qui conduit irrémédiablement à des pertes d'information en bordures des traits, essentiellement sur les zones extrémales.

Dans **l'étape *3*** l'utilisation simultanée des rayons extraits à partir de la transformée en distance et de l'axe médian calculé à partir de la diffusion du gradient va nous permettre de préparer le suivi du tracé. Nous avons fait une proposition qui améliore la méthode de suivi de tracé proposée par *(Xu et al. [2007])* en utilisant en chaque point, la direction de la diffusion de gradient qui remplace les deux directions complémentaires (géométrique et de l'intensité) utilisées dans la méthode de Xu. La direction de diffusion fournit une meilleure précision par rapport à la variation locale de l'intensité causée par la dégradation présente sur le niveau de

l'encre ou du papier. Cette direction sera utilisée durant le suivi du tracé pour trouver les points qui vont appartenir à l'axe médian.

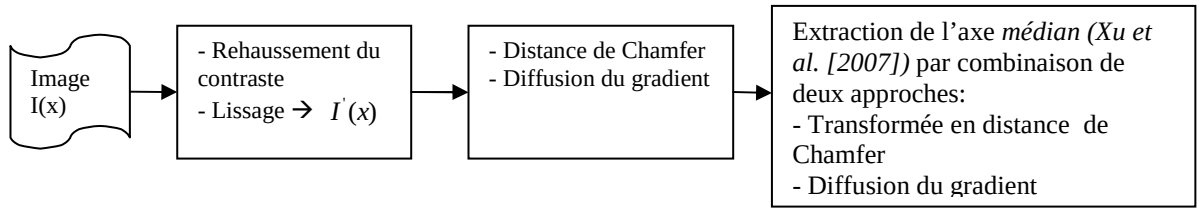


Figure 3.6. Schéma général de principe d'extraction de l'axe médian : du rehaussement au découpage

3.1.1 Rehaussement du contraste

Cette étape de rehaussement de contraste est appliquée à l'image en niveaux de gris pour renforcer et améliorer les structures. Pour cette application nous avons utilisé le filtre d'amélioration de Frangi que nous décrivons en détail dans cette partie. Cette méthode est basée sur les valeurs propres de la matrice Hessienne (matrice des dérivés partielles secondes) évaluées en chaque pixel de l'image et considérant différentes échelles permettant de considérer l'épaisseur maximum d'un trait. Elle possède la propriété d'être robuste aux dégradations des traits. Nous avons donc choisi de l'appliquer aux images de manuscrits bruités du patrimoine médiéval. La méthode de Frangi est appliquée sur l'image selon le mécanisme général suivant:

Dans la première étape on construit les dérivées secondes de gaussienne de l'image I pour une échelle σ :

$$\frac{\partial^2 I}{\partial x^2} = \frac{1}{2 \cdot \pi \cdot \sigma^4} \cdot \frac{X^2}{(\sigma^2 - 1)} \cdot e^{-(X^2+Y^2)/2\times\sigma^2} \quad \textbf{(3.1)}$$

$$\frac{\partial^2 I}{\partial x \partial y} = \frac{1}{2 \cdot \pi \cdot \sigma^6} \cdot X \cdot Y \cdot e^{-(X^2+Y^2)/2\times\sigma^2} \quad \textbf{(3.2)}$$

$$\frac{\partial^2 I}{\partial y^2} = \left(\frac{\partial^2 I}{\partial y^2} \right)^T \quad \textbf{(3.3)}$$

Dans la seconde étape on construit la matrice Hessienne *H(I)* de l'image *I* à partir des dérivées secondes

$$H(I) = \begin{pmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} \end{pmatrix} \quad \textbf{(3.4)}$$

Les valeurs propres λ_1 et λ_2 peuvent être utilisées pour classifier les pixels des images comme étant dans des structures linéaires ou épaisses (blobs) (tableau 3.5).

Tableau 3.5. Modèles possibles en *2D* pour les valeurs propres λ_1 et λ_2 de la matrice Hessienne avec *P* = valeur petite, *E* = valeur élevée, +/- indiquent le signe des valeurs propres selon la condition $|\lambda_1| \leq |\lambda_2|$

	λ_1	λ_2	**Structure**
Modèle 1	*P*	*E*-	Structure linéraire (blanche sur fond noire)
Modèle 2	*P*	*E*+	Structure linéaire (noire sur fond blanc)
Modèle 3	*P*-	*E*-	Structure Blob (blanche sur fond noire)
Modèle 4	*P*+	*P*+	Structure Blob (noire sur fond blanc)

Donc tout pixel qui ne répond pas au deuxième modèle sera écarté. Pour une structure linéaire, la plus petite valeur propre λ_1 , correspond au vecteur propre u_1 qui est dans la direction tangentielle de la structure. Inversement à la direction normale associée à la valeur propre λ_2 , où l'intensité est marquée par une variation brutale entre le trait et l'arrière plan, l'intensité ne varie pas dans la direction tangentielle. Nous définissons ainsi une mesure de *Blobness* $R_\beta = |\lambda_1 / \lambda_2|$, basée sur ce rapport entre les vecteurs propres u_1 et u_2.

De même pour une structure de trait noir bien définie nous constatons en chaque pixel une valeur de λ_2 élevée, nous ajoutons alors une deuxième mesure $S = \sqrt{|\lambda_1|^2 + |\lambda_2|^2}$ pour identifier ces structures. Ces deux mesures sont combinées pour calculer la mesure de *Vesselness* représentée par la fonction v_0 qui se calcule avec une valeur de sigma σ fixée pour un pixel x de la façon suivante :

$$v_0(s) = \begin{cases} 0 & \lambda_2 \prec 0 \\ \exp\left(-\frac{R_{\beta^2}}{2\beta^2}\right)\left(1-\exp\left(-\frac{S^2}{2c^2}\right)\right) & \text{autrement} \end{cases} \quad \textbf{(3.5)}$$

La fonction v_0 est constituée des parties normalisées suivantes dans l'intervalle *[0, 1]:*

- R_β doit avoir une faible valeur
- S doit avoir une grande valeur

Quand R_β augmente, la valeur de *exp()* diminue de *1* vers *0*, et quand *S* augmente la valeur de $\left(1-\exp\left(-\frac{S^2}{2c^2}\right)\right)$ augmente de *0* vers *1*, avec $\beta = 0{,}5$ et $c = 0{,}5$ deux valeurs de seuils qui contrôlent la sensibilité de la mesure R_β.

La valeur résultante de v_0 au niveau d'un pixel *x* est calculée comme étant la réponse maximale sur toutes les valeurs de σ :

$v_0(x) = \max(v_0(x,\sigma))$ pour σ appartenant à l'intervalle $\langle \sigma_{\min}, \sigma_{\max} \rangle$. L'échelle σ pour laquelle *v* atteint son maximum détermine la taille de la structure.

L'algorithme de rehaussement de contraste de Frangi est alors défini comme suit:

Algorithme de rehaussement de contraste de Frangi
Pour chaque σ faire : Pour chaque pixel faire : • Calculer les valeurs propres λ_1 et λ_2 de la matrice Hessienne pour chaque valeur de σ . • Ordonner et étiqueter les valeurs propres de sorte que $\lvert\lambda_1\rvert \le \lvert\lambda_2\rvert$ • Pour les traits d'écriture qui représentent des structures noires sur un fond blanc on doit avoir $\lvert\lambda_1\rvert \approx 0, \lvert\lambda_1\rvert \le \lvert\lambda_2\rvert$ • Calculer la valeur scalaire de la fonction v_0

3.1.2 Principe de l'extraction de l'axe médian par la diffusion du gradient

3.1.2.1 Principe de la diffusion du gradient dans une image, exploitation pour l'extraction de squelette

Dans cette section nous expliquons le principe de la méthode de diffusion du gradient utilisée pour l'extraction de l'axe médian. Dans nos travaux sur les manuscrits médiévaux nous traitons

des images bruitées et de l'encre vieillissant, c'est pour cela que les opérations morphologiques ne sont pas intéressantes. Nous avons mentionné dans la section 3.2.2 que les algorithmes nécessitant une étape de binarisation conduiront à une perte d'information. Par ailleurs, l'utilisation de méthodes heuristiques nous obligerait à régler de nombreux paramètres. Se baser sur les contours pour extraire l'axe médian n'est pas une bonne option car nous travaillons sur des documents dégradés et nous pouvons rencontrer des discontinuités de contour. Pour ces raisons nous avons décidé d'utiliser la diffusion de gradient qui nous permet de :

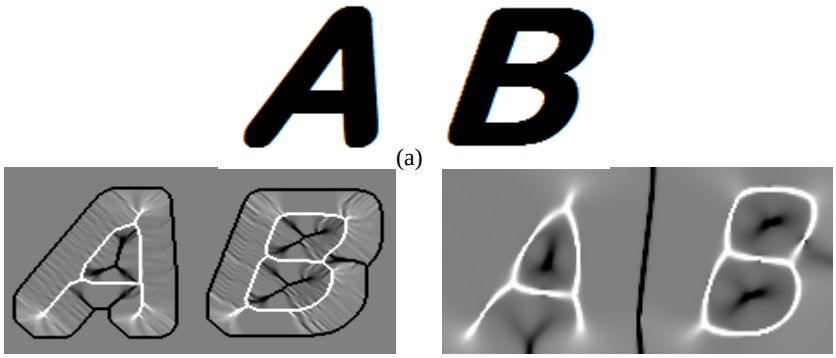
- Dépasser les heuristiques, paramètres, masques et sélection de points.
- D'initialiser l'algorithme de diffusion de gradient sans avoir besoin de points de départ qui appartiennent à un objet (dans notre cas un trait).
- D'avoir la possibilité de travailler sur des images de manuscrits médiévaux bruitées.
- D'appliquer la diffusion du gradient directement sur l'image en niveaux de gris ou bien en couleur sans avoir besoin de passer par une étape de binarisation. Dans la méthode de (Lebourgeois et E*mptoz, [2007])* le gradient est appliqué sur l'image entière où on calcule l'orientation du gradient. L'idée principale de cet algorithme est basée sur la correction de la direction du gradient en régularisant le champ de vecteur du gradient qui est présenté par l'équation suivante :

$$\nabla I^{n+1} = \frac{1}{|N(p)|} \sum_{h \in N(p)} \nabla I^{n}(h), et \quad \nabla I = \begin{pmatrix} I_x \\ I_y \end{pmatrix} \tag{3.6}$$

On note ∇I^n le champ de vecteur de gradient régularisé à l'ordre *n*.

La régularisation peut être considérée comme un processus itératif de lissage des vecteurs gradients en utilisant *N(p)*, les *8*-voisins connectés de chaque point p, avec un noyau uniforme. La régularisation excessive du champ de vecteur de gradient va progressivement diverger. Pour maintenir la stabilité du champ de gradient par la régularisation du vecteur, nous avons besoin, soit de réduire le nombre d'itérations ou d'utiliser un algorithme de feu de brousse ou d'une fonction d'arrêt pour bloquer la diffusion avant d'atteindre la perte de stabilité du champ. Nous avons choisi de contrôler la régularisation en arrêtant le flux du gradient à partir de l'équation précédente, quand le nombre d'itérations n pour un gradient non nul $\nabla I^n \neq 0$ atteint une limite l définie par l'utilisateur. Si l = 1, nous avons alors un algorithme de feu de brousse classique qui fournit un axe médian avec des branches parasites. Pour une plus grande valeur de l, on obtient un axe médian lisse, robuste au bruit et adapté pour les images réelles (Figure 3.7).

Le paramètre de lissage l peut être augmenté pour analyser les images bruitées sans atteindre la perte de stabilité du champ de gradient en utilisant la régularisation avec le vecteur de gradient.



(a)

(b) (c)

Figure 3.7. (a) image originale, (b) algorithme feu de brousse avec l = 1 et $\theta(\nabla I^{n=8})$, (c) l =40 et $\theta(\nabla I^{n=82})$

La figure 3.8 montre le champ de gradient sur une image bruitée (haut et bas à gauche), et le champ de gradient qui est généré après la régularisation (haut et bas à droite). Durant ce processus de régularisation, les vecteurs de gradient ayant une plus grande amplitude sont réorientés correctement jusqu'à la converge et l'arrêt des itérations. $\theta(\nabla I) = arctg\left(\frac{I_y}{I_x}\right)$ présente l'orientation du gradient et n le nombre d'itérations.

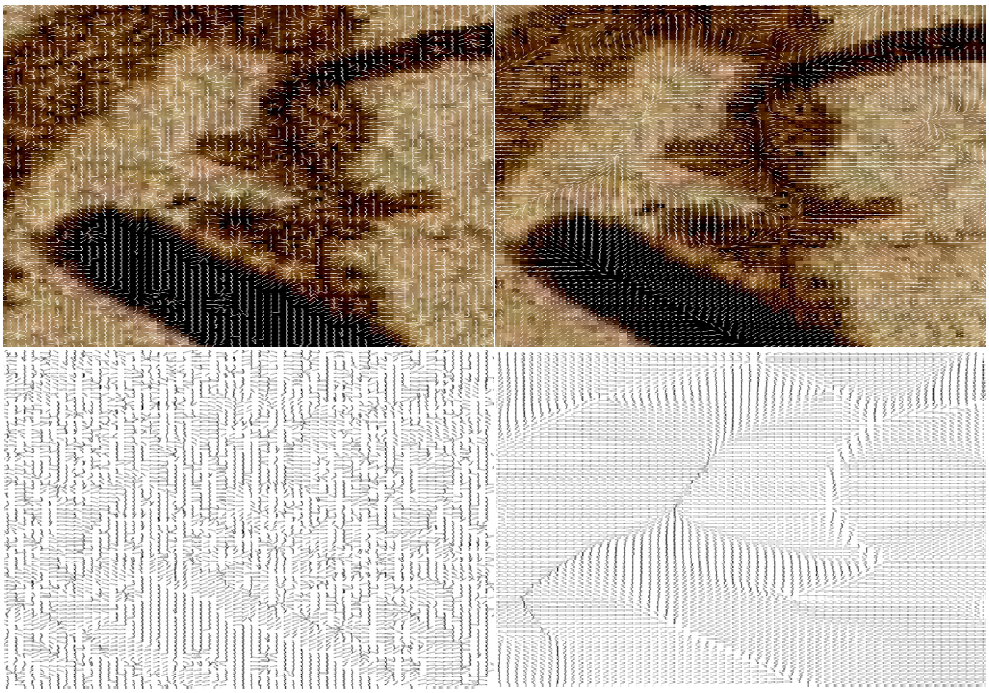
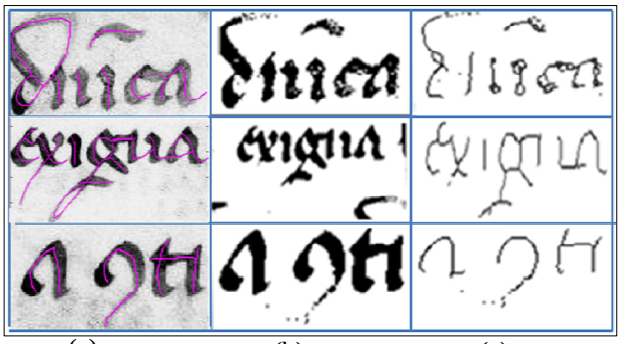


Figure 3.8. Régularisation et diffusion du gradient sur un extrait de texte ancien

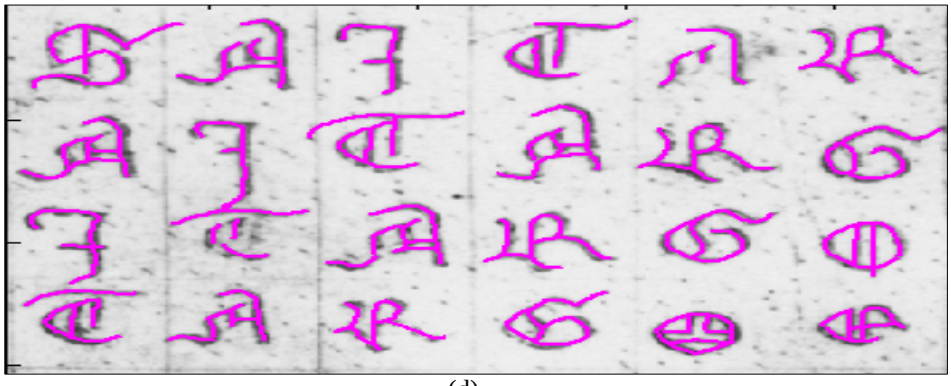
L'orientation corrigée va être utilisée pour calculer la carte de force du squelette (CFS). L'axe médian est extrait comme suit : la différence d'angle maximal de l'orientation de deux paires symétriques et adjacentes de vecteurs de gradient dans une fenêtre de taille 3×3 est basée sur l'équation suivante:

$$SS(p) = \max_{h,k\in N(p)} \{|\theta(h) - \theta(k)|\} \quad \textbf{(3.7)}$$

$\theta(h)$ et $\theta(k)$ présentent les orientations du gradient des paires symétriques ou adjacentes de vecteurs de gradient dans un voisinage connecté de taille 8. Si *SS (p) = 180°* alors le point *p* appartient à l'axe médian. Sinon si *SS (p) = 45°* alors le point est localisé sur une forme triangulaire ou bien des branches parasites du squelette. La figure 3.9 montre le résultat de l'extraction de l'axe médian par notre méthode. Basée sur la méthode de Xu, il est nettement meilleur que celui obtenu par la méthode classique de Zhang qui est la technique communément employée pour extraire rapidement un squelette sur une image binaire *(Zhang et Suen, [1984])*. On remarque que notre méthode a bien détecté l'axe médian, même dans les situations où l'encre était dégradée ou très claire (extrémités de lettres, boucles présentes dans les jambages des lettres…).



(a) (b) (c)



(d)

Figure 3.9. Extraction de l'axe médian par, (a) notre méthode, (c) la méthode de Zhang. (b) binarisation par la méthode de Sauvola, (d) résultat de notre méthode sur un manuscrit dégradé (encre claire, bruit, trait effacé)

3.1.2.2 Méthode de suivi de squelette pour l'extraction de l'axe médian

Les étapes de notre méthode de suivi et de détection de l'axe médian sont résumées dans l'algorithme suivant. Elles sont inspirées des travaux de Xu dans *(Xu et al. [2007])* :

Algorithme de suivi de tracé inspiré de la méthode de Xu dans *(Xu et al. [2007])*

a) Initialisation

- Détecter les points de départ issus de la diffusion (un point est considéré comme point de départ si son intensité lumineuse est maximale par rapport à ses voisins).
- Extraire le rayon R_{k+1} de chaque point de départ P_k (calculé à partir de la distance de Chamfer pondérée).
- Commencer à partir du premier point de départ rencontré.
- Déterminer le point suivant P_{k+1}^0 dans d_k pixels ($d_k = 1$, appelée « look ahead distance ») du point de départ P_k dans la direction du gradient ψ_{k+1}.
- Calculer la direction ψ_{k+1}^0 (direction de la diffusion du gradient) du point P_{k+1}^0.

b) Détermination du point suivant et ajustement de sa position

- Tracer un segment g_{k+1} avec une épaisseur égale à $2 \times R_{k+1}^0$ du point P_{k+1}^0.
- Mettre à jour la direction ψ_{k+1}^1 pour déterminer le point suivant P_{k+1}^1 qui aura une valeur de rayon qui est déjà calculé à partir de la distance de Chamfer pondérée.
- Procéder de la même façon par calcul du profil d'intensité pour trouver le point focal P_{k+1}, point de convergence au centre du trait et calculer sa direction ψ_{k+1}.
- Marquer ce point comme un point visité, de cette façon il ne sera pas visité une autre fois par notre suivi de tracé.
- Réitérer l'étape (b) le long du tracé jusqu'à atteindre un critère d'arrêt, dans ce cas choisir le point de départ suivant P_k et reprendre l'étape (b).

c) Critères d'arrêt

- Dans le cas où on rencontre un point de bifurcation (intersection de deux traits) et que ce point a déjà été visité, on arrête le suivi du tracé.
- Si on arrive à un point marqué comme point de départ, il sera enlevé de la liste des points de départ.

- Si l'un des critères d'arrêt est vérifié et la liste des points de départ est vide alors on arrête l'algorithme de suivi du tracé.

La figure 3.10, illustre l'algorithme de suivi de tracé. Les segments ($g_k, g_{k+1}, \tilde{g}_{k+1}$) qui représentent les profils d'intensité, sont égales respectivement à ($2\times R_k, 2\times R^0_{k+1}, 2\times R^1_{k+1}$) qui représentent les rayons calculés à partir de la distance de Chamfer pondérée (Meyer et Maragos, [1999]) sur les points ($P_k, P^0_{k+1}, P^1_{k+1}$). Les directions ($\hat{\psi}_k, \hat{\psi}^0_{k+1}, \hat{\psi}_{k+1}, \hat{\psi}^1_{k+1}$) sont calculées à partir de la diffusion de gradient respectivement sur les points ($P_k, P^0_{k+1}, P_{k+1}, P^1_{k+1}$). P_{k+1} est considéré comme le point focal qui va appartenir à l'axe médian avec une direction $\hat{\psi}_{k+1}$.

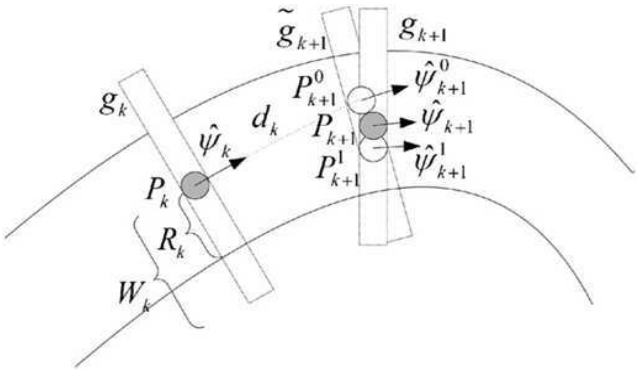


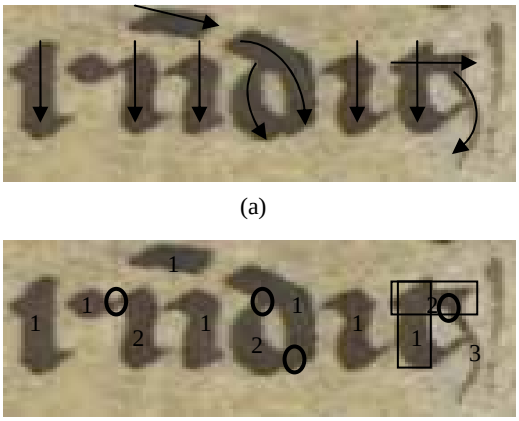
Figure 3.10. Illustration de l'algorithme de suivi du tracé et extraction de l'axe médian

Dans la section suivante nous décrivons notre méthode de décomposition des manuscrits en graphèmes à partir de l'extraction de l'axe médian et du suivi de tracé précédemment décrit.

4 Décomposition en traits

En paléographie l'analyse du ductus révèle des informations sur l'époque du manuscrit et sur le style utilisé. L'analyse du ductus dans les manuscrits gothiques montre que les traits se font de gauche à droite et de haut en bas figure 3.11(a). Les gestes de rebroussement (retour en arrière de la plume) sont impossibles. De tels mouvements sont considérés comme des mouvements incompatibles avec la nature des plumes et du papier à l'époque du Moyen Age. L'épaisseur de l'encre joue aussi un rôle important qui indique la fin d'un trait et le début d'un autre, voir figure 3.11 (b). Sur cette figure, on constate que la lettre *'i'*, première lettre à gauche se fait en un seul geste de la plume. La lettre *'d'* se construit en deux mouvements. Une variation importante dans l'épaisseur d'encre, marque qu'un nouveau trait entre dans la formation de la lettre. Les points de décomposition sont indiqués par des cercles. Les croisements illustrés par des rectangles sur la figure 3.11(b) indiquent une zone de superposition de deux traits qui se

traduira par un point de décomposition. Notons également que ces règles s'appliquent de la même façon sur les manuscrits de l'époque carolingienne.



(a)

(b)

Figure 3.11. Cette figure montre d'une part (a) les règles de formations des traits et d'autre part, (b) les règles de décomposition

La décomposition se base sur les règles paléographiques décrites précédemment et fonctionne de la façon suivante : l'axe médian obtenu dans la section 3 est utilisée pour découper l'écriture manuscrite en graphèmes. Entre chaque point de départ et d'arrêt d'un trait, tous les points de l'axe médian sont enregistrés dans une liste comprenant leur direction (direction du gradient) et leur épaisseur *(2×R)*. *R* représente la demi-épaisseur du trait calculée à partir de la distance de Chamfer pondérée.

Les points d'épaisseur minimale (minimum local) sont ensuite marqués et proposés comme point de découpage, comme cela est effectivement le cas dans la formation d'un trait (Figure 3.12).

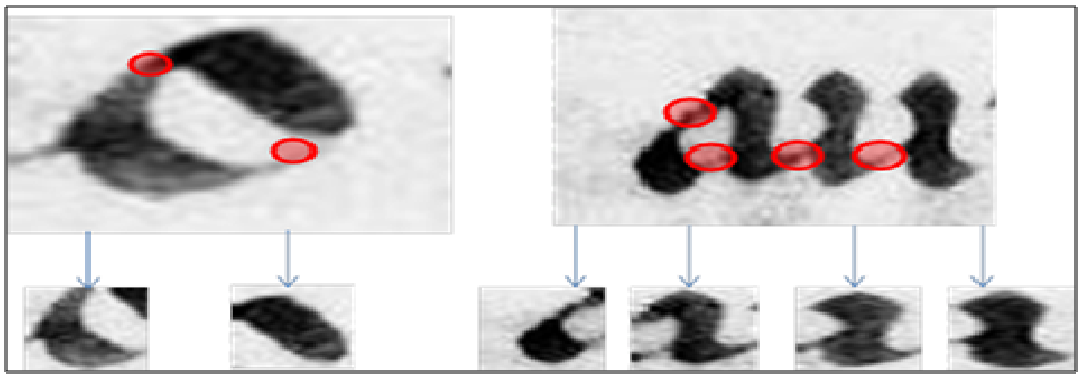


Figure 3.12. Points de décomposition aux points minimaux locaux des traits

La décomposition présentée dans la figure 3.13(a) montre que les lettres sont constituées de fragments adjacents rattachés aux points d'épaisseur minimale. Certains de ces points sont supposés correspondre à des points de poser et de lever de plume. Comme le montrent les courbes de la figure 3.13(b), sur la lettre *«C»*, on ne localise qu'un seul point de découpage alors que pour la lettre *«O»* il y en a deux. La figure 3.13(c) montre des exemples de décompositions se basant sur le minimum local.

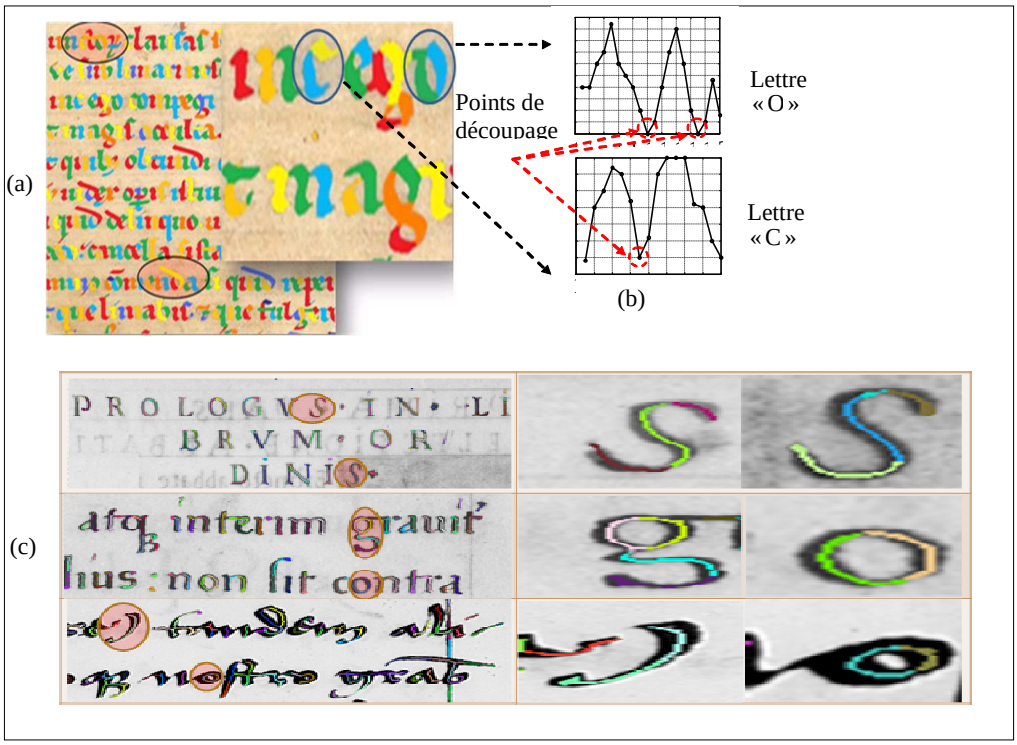


Figure 3.13. (a) Exemple de décomposition des traits en graphèmes par notre méthode, (b) Courbes des épaisseurs des points de l'axe médian des lettres « O » et « C», et points de découpage en graphèmes, (c) résultats de découpage sur d'autres styles d'écritures

Si l'algorithme rencontre un croisement de traits, la décomposition va dépendre de l'ordre dans lequel les traits sont positionnés.

Nous commençons tout d'abord par définir le cas d'un croisement. Pour cela nous introduisons un paramètre de niveau de gris normalisé τ pour chaque point de l'axe médian. Pour un point P_{k+1}, τ_{k+1} est défini par :

$$\tau_{k+1} = \frac{I_{\max}(k+1) - I_{\min}(k+1)}{I_{\max}(k+1)} \qquad \textbf{(3.8)}$$

Où *Imax* et *Imin* représentent l'intensité maximale et minimale sur le profil d'intensité g_{k+1}, segment perpendiculaire à la direction du trait au point P_{k+1}. La taille du segment est initialisée à R_k

pendant le calcul de τ_{k+1}. Dans le cas d'un croisement de trait, les intensités minimales et maximales diminuent. La diminution de l'intensité minimale est généralement plus importante que celle de l'intensité maximale, résultant en un changement radical dans la valeur de τ ; tandis que dans d'autres cas, τ reste à peu près le même. Sur la base de cette observation, nous jugeons la situation de croisement lorsque : $\tau_{k+1} \geq \beta_{\tau 0}$. $\tau 0$ représente la valeur moyenne de τ des trois points précédents sur l'axe médian si ils ne présentent pas une situation de croisement. La moyenne est calculée pour réduire la perturbation de bruit. β est un facteur de pondérations et il est fixé à *1,5* dans notre algorithme.

La figure suivante nous montre une situation de croisement, l'ordre de suivi des traits est représenté par les deux flèches (1) et (2), voir figure 3.14 (a). Le second suivi de tracé va couper le premier tracé en deux parties, pour cela on va extraire 3 traits qui sont : les deux parties du premier trait et le trait du deuxième suivi. L'ordre de suivi des traits n'est pas encore contrôlé par notre algorithme, une suggestion actuellement à l'étude consiste à envisager des possibilités de découpage en commençant tout d'abord par les traits horizontaux puis en continuant par les traits verticaux ou inversement. La figure 3.14 (b) montre un cas de croisement où on observe une diminution de l'intensité lumineuse en son centre. Enfin la figure 3.14 (c) montre le cas de croisement de la lettre '*x*' qui a été décomposée en trois segments, la deuxième ligne bleue (figure 3.14 (c) gauche) a coupé la première en deux segments. Identiquement sur la figure 3.14 (c) droite, le segment vert clair découpe le segment en deux parties.

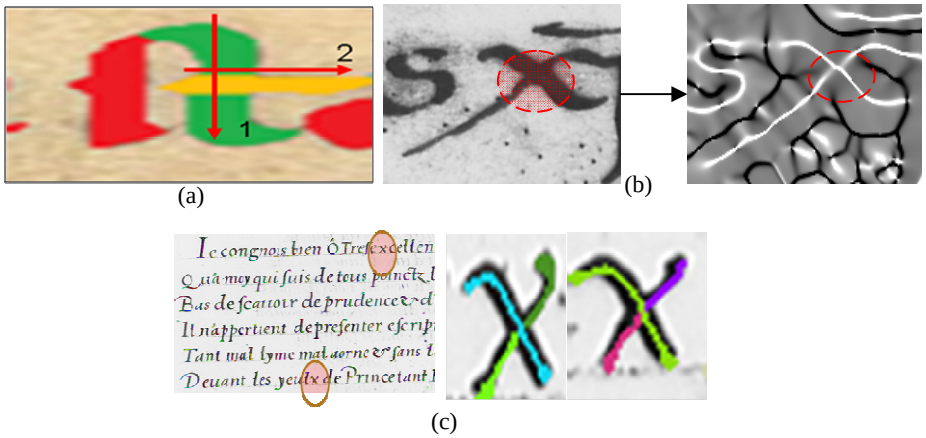
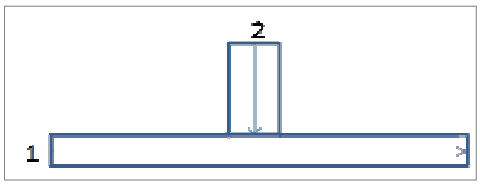
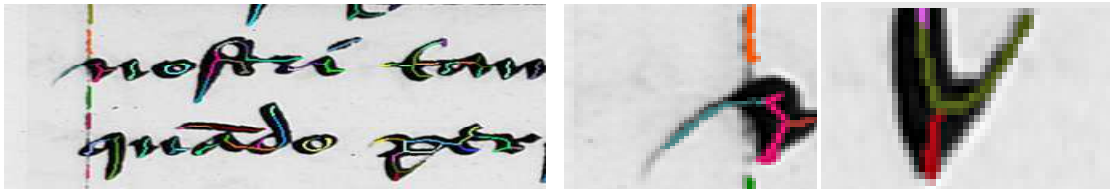


Figure 3.14. (a) ordre des traits, (b) point de décomposition indiquant une variation dans l'intensité du gradient, (c) exemples de décomposition de deux lettres « *X* » aux points de croisement produisant trois segments

Dans l'hypothèse où l'on rencontre un trait déjà visité lors du suivi du segment principal, le suivi du second trait est automatiquement arrêté, et nous considérons que nous avons un point de découpage. Nous appelons ces situations des cas de jonction, voir figure 3.15.



(a)



(b)

Figure 3.15. Illustration des points de jonction simple. (a) ordre de visite des traits, (b) exemple d'intersection sur des manuscrits médiévaux

4.1 Étude de la stabilité de la décomposition

Dans cette section, nous avons choisi d'étudier la stabilité de la décomposition des lettres en graphèmes selon notre méthode de suivi de traits et l'exploitation des règles d'exécution des traits produits par les copistes lors de la conception des textes. Signalons avant toute chose que par hypothèse, les formes manuscrites produites sur les documents de notre étude sont très stables et « ressemblantes » : en effet, les lettres sont formées selon des règles paléographiques strictes, ce qui leur confère des propriétés visuelles très similaires. Sans cette hypothèse, il est impossible d'envisager d'estimer une stabilité de décomposition des lettres si celles-ci sont toutes exécutées de façon différente.

L'étude la stabilité de la décomposition a été faite visuellement sur un ensemble de lettres, les plus utilisées dans les manuscrits de la base IRHT : *a, b, d, e, f, o* (figure 3.16).

Figure 3.16. Exemples des lettres les plus utilisées dans les manuscrits

Pour chaque lettre, nous avons sélectionné au hasard 8 manuscrits. A partir de chaque manuscrit, nous avons extrait 5 occurrences de chaque lettre. Nous avons donc au total et pour

chaque lettre *40* occurrences (voir figure 3.17) ce qui conduit à un total de *240* occurrences pour l'ensemble des lettres.

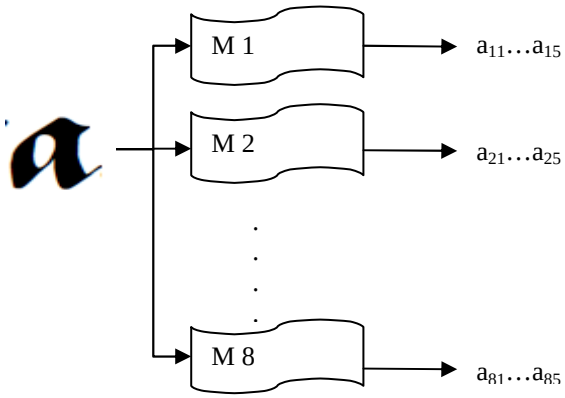


Figure 3.17. Schéma de formation des occurrences de la lettre « a » sur huit manuscrits

Pour calculer la stabilité de décomposition d'une lettre pour chaque manuscrit *M* nous utilisons la formule suivante :

$$S_M = \frac{(C-1)}{(T-1)} \qquad \textbf{(3.9)}$$

Où *C* représente le nombre de variations de décomposition d'une même lettre au sein du manuscrit (par exemple, sur les 5 occurrences au total, si on constate 5 décompositions différentes, ce qui correspond au cas le plus défavorable, alors C prendra la valeur de 5). La composante T est fixée à *5*. Elle correspond au nombre total d'occurrences retenues pour chaque lettre. Par exemple, pour la lettre « a » du manuscrit *M* : *a1*, *a2* et *a4* ont visuellement la même décomposition, mais *a3* et *a5* sont décomposées en fragments différents (figure 3.18).

Figure 3.18. Exemple d'occurrences de la lettre « a » et leur segmentation pour le manuscrit 1 avec *C = 3* et S_{M1} *= 0,5*

Pour la chaque lettre, la stabilité globale sur l'ensemble des *8* manuscrits étudiés est définie par $S = S_{M1} + ... + S_{Mn}/8$. Plus *S* est proche de *0*, plus la décomposition est stable et plus S est proche de 1 plus la décomposition est instable.

La stabilité de la décomposition dépend très fortement des paramètres fixés dans l'étape de diffusion utilisée pour extraire l'axe médian qui sont : $\sigma_{Gaussienne}$, l ainsi que du paramètre *n*

représentant le nombre d'itérations appliquées pour faire converger les directions de gradients et qui est lié à la présence de distorsions, de bruits, à la complexité des formes et l'existence de dégradations locales de l'encre qui peuvent coexister dans les manuscrits médiévaux.

Nous montrons dans la figure 3.19 les variations des valeurs de la stabilité globale *S* pour deux valeurs de paramètres de diffusion. Pour le premier cas : $\sigma_{Gaussienne} = 1{,}2$, $l = 8$ et $n = 8$ et pour le second : $\sigma_{Gaussienne} = 0{,}6$, $l = 5$ et $n = 5$. Nous remarquons dans la figure 19 que le premier cas donne une meilleure stabilité sur toutes les lettres. Pour la lettre *« o »* qui n'est pas considérée comme une forme complexe nous avons des valeurs de *S* plus petites que dans les autres cas. En revanche, la plus mauvaise stabilité est produite pour la lettre *« f »*.

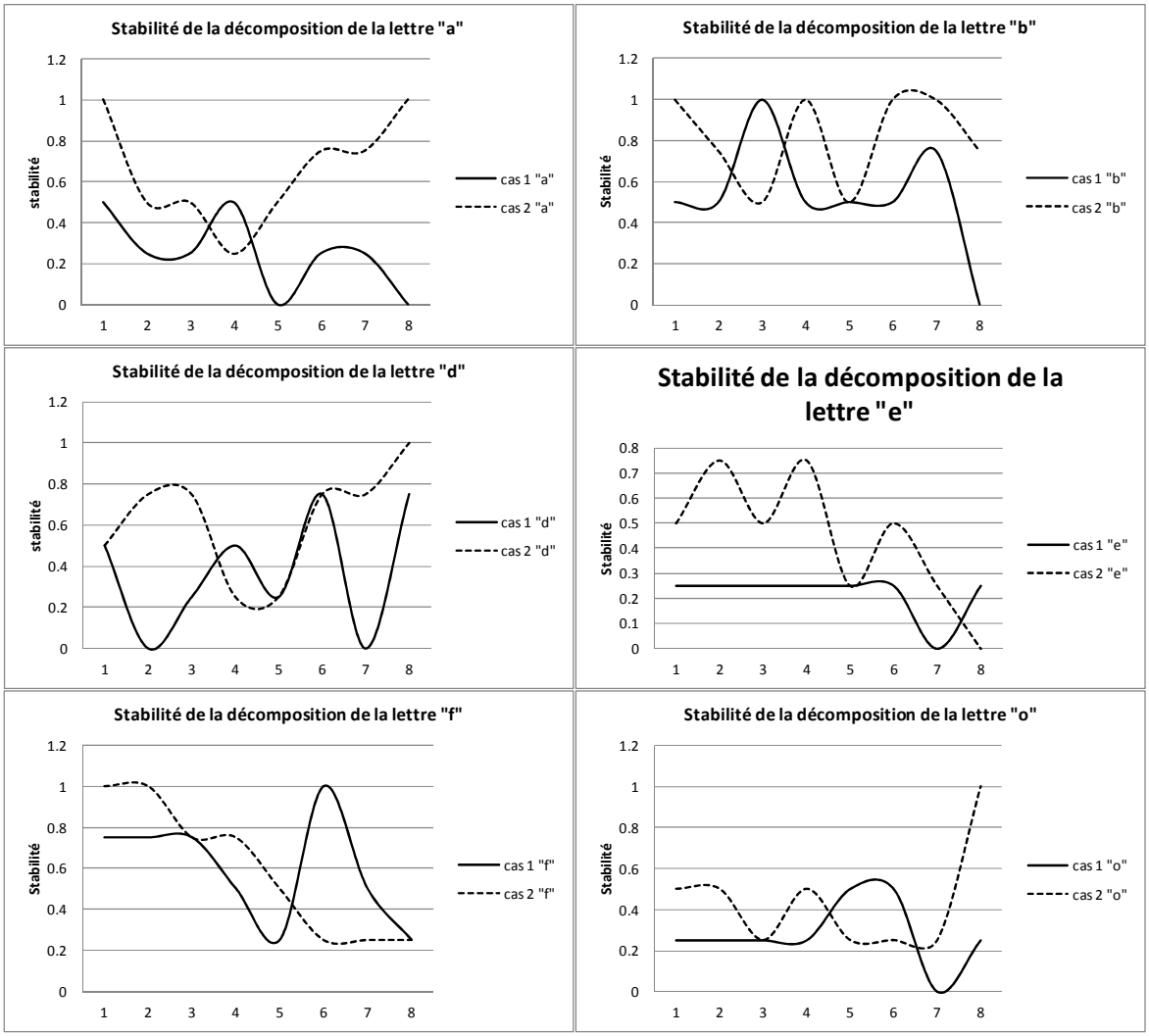


Figure 3.19. Valeurs de *S* pour les lettres « *a, b, d, e, f, o* », L'axe des abscisses présente les 8 manuscrits retenus pour mener les tests

Parallèlement aux valeurs de *S* pour l'ensemble des lettres, nous indiquons à la figure 3.20 les valeurs de *S* cumulés exprimées pour l'ensemble des lettres dans les deux cas de configuration de l'algorithme de diffusion. Une bonne décomposition indique une faible variation (voir une variation nulle) dans la décomposition des lettres d'un manuscrit.

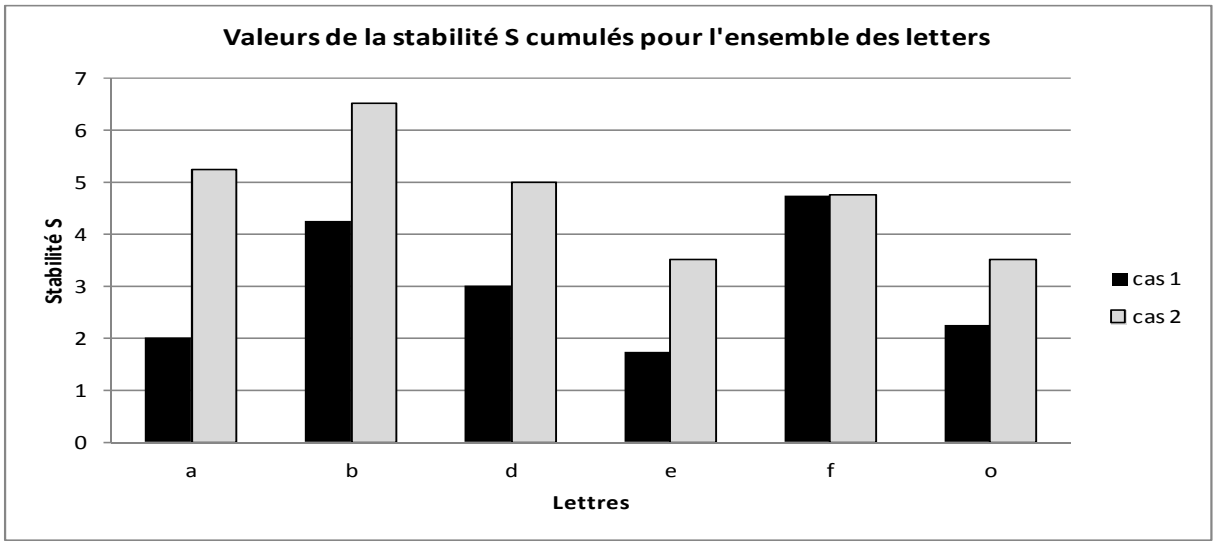


Figure 3.20. Valeurs de la stabilité *S* pour les lettres « a, b, d, e, f, o ». Pour toutes les lettres, le meilleur taux de décomposition est donné par le premier cas (cas 1)

Ces résultats indiquent que le paramétrage de la diffusion responsable de la détermination de l'axe médian est très important et qu'il constitue la partie sensible de la méthode.

5 Conclusion

Dans ce chapitre nous avons présenté notre méthode de suivi de tracé et de décomposition des traits en graphèmes. Nous avons contourné les problèmes rencontrés dans les méthodes locales telles que la binarisation qui est souvent une étape indispensable pour l'extraction de squelette et qui est considérée comme un inconvénient lorsqu'elle est appliquée sur les manuscrits anciens de mauvaise qualité. Nous avons montré que la partie de prétraitement du manuscrit permettant d'obtenir une bonne décomposition était fondamentale et difficile à concevoir en raison des très nombreuses contraintes rencontrées sur les manuscrits anciens, essentiellement liées à la présence de bruit distribué sur l'ensemble des images (fond et formes).

Notre proposition de décomposition en graphèmes constitue également une réponse faite aux experts paléographes dans sa capacité à mettre en lumière sur les images une décomposition compatible avec les règles d'exécution des formes écrites originellement.

La décomposition souffre encore actuellement de certaines limites de stabilité lorsque les épaisseurs de traits ne sont pas intégralement contrôlables ou lorsque les situations de croisement sont floues et difficiles à bien repérer. Les variations internes de luminosité (liées

aux variations des niveaux de gris des encres) peuvent en effet conduire à des effets de sur-segmentation indésirables. Ces différents points font partie des perspectives de nos travaux visant une meilleure adéquation entre notre processus de décomposition et le processus réel d'exécution des traits.

Le chapitre suivant est dédié à la caractérisation des graphèmes segmentés et à notre approche de la pondération de ces caractéristiques assurant le meilleur partitionnement des formes pour la construction des dictionnaires, qui représentent les signatures uniques des manuscrits.

Chapitre 4 : Caractérisation des graphèmes et construction de dictionnaire de formes

Résumé: *Nous présentons dans ce chapitre une nouvelle approche de caractérisation de l'écriture par la génération de dictionnaires de formes basés sur un mécanisme de sélection et de pondération de caractéristiques par algorithme génétique et clustering par coloration minimale de graphe. L'approche proposée est basée sur la décomposition du manuscrit en graphèmes. Par rapport à d'autres techniques de réduction de dimension comme l'ACP nous proposons une classification non-supervisée des graphèmes à partir de l'évaluation des dissimilarités entre graphèmes finalement rassemblés en dictionnaire de formes. Un processus de coloration minimale de graphe est appliqué pour la catégorisation des graphèmes et permet de faire converger l'algorithme génétique vers des solutions optimales. Dans ce chapitre, nous montrons l'intérêt du couplage entre l'estimation de poids associés à chaque descripteur (déduits de l'algorithme génétique) et les résultats de clustering des graphèmes par coloration minimale de graphe.*

Mots clés: *algorithmes génétiques, sélection et pondération de caractéristiques, dictionnaires de formes, fitness.*

1 Introduction

Après avoir segmenté les manuscrits en graphèmes (cf. Chapitre *3*), ces derniers sont caractérisés par un ensemble de *59* caractéristiques et classifiés en groupes homogènes formant des dictionnaires de formes représentatifs de chaque manuscrit dans la base.

Pour améliorer la précision de la classification, des techniques comme la *pondération de caractéristiques* où chaque caractéristique est multipliée par une valeur proportionnelle à sa capacité de discrimination et la *sélection de caractéristiques* dédiée à la sélection d'un sous-ensemble de caractéristiques pertinentes qui ignore toutes les caractéristiques non significatives pour chaque type de manuscrit. Ces deux techniques ont trois objectifs :

1. Réduire le coût d'extraction des caractéristiques
2. Améliorer la précision de la classification
3. Améliorer la fiabilité de l'estimation de la performance.

Un autre avantage important lié à la sélection de caractéristiques est la possibilité d'introduire l'expert humain (paléographe) dans la boucle de sélection en rendant plus interactif

le processus de sélection et en permettant à l'expert d'introduire des connaissances directement dans le processus de caractérisation. Dans ce chapitre, nous présentons le principe de sélection et de pondération des caractéristiques portant sur les algorithmes génétiques (AG). Des poids génériques accompagnant chaque caractéristique sont déduits du processus de sélection de caractéristiques en utilisant un modèle d'apprentissage et de test. Nous montrons comment la combinaison de la sélection et de la pondération de caractéristiques peut améliorer la performance de la classification. Dans ce chapitre nous allons présenter en premier un état de l'art comportant une partie sur les différentes familles de sélection de caractéristiques, puis sur les méthodes de sélection et pondération de caractéristiques par algorithmes génétiques utilisées dans le domaine d'analyse des manuscrits. Puis, nous illustrerons nos deux approches de sélection et de pondération de caractéristiques en détaillant le choix des descripteurs retenus. Nous présenterons dans le détail le mécanisme de génération de poids automatiques et l'exploiterons sur un exemple concret.

2 Travaux antérieurs

2.1 Les familles de méthodes de sélection de caractéristiques

La sélection de caractéristiques est une technique permettant de choisir les caractéristiques les plus intéressantes ou discriminantes afin de réaliser par exemple une tâche de reconnaissance.

De manière générale, la sélection de caractéristiques a plusieurs grands intérêts :

- Réduction de la dimension de l'espace des caractéristiques, surtout lorsque le nombre de variables atteint plusieurs milliers. Les algorithmes d'apprentissage et de classification ne peuvent pas effectuer leur traitement dans des temps raisonnables.
- Simplification du modèle de représentation des données pour le classifieur. La réduction de la dimension de l'espace des caractéristiques permet alors de réduire le nombre de paramètres nécessaires à la description du modèle.
- Amélioration des performances de la classification, sa vitesse et son pouvoir de généralisation.
- Augmentation de la compréhensibilité des données : on voit mieux quels sont les processus qui leur ont donné naissance et leur rôle dans le processus de traitement (classification).
- L'élimination des variables indépendantes de la classification.
- L'élimination des variables redondantes.

La figure 4.1, illustre la structure générale des méthodes de sélection de caractéristiques. La sélection de caractéristiques consiste à générer des sous-ensembles *Y* en parcourant l'espace des sous-ensembles qui sont évalués jusqu'à ce qu'un certain critère soit satisfait.

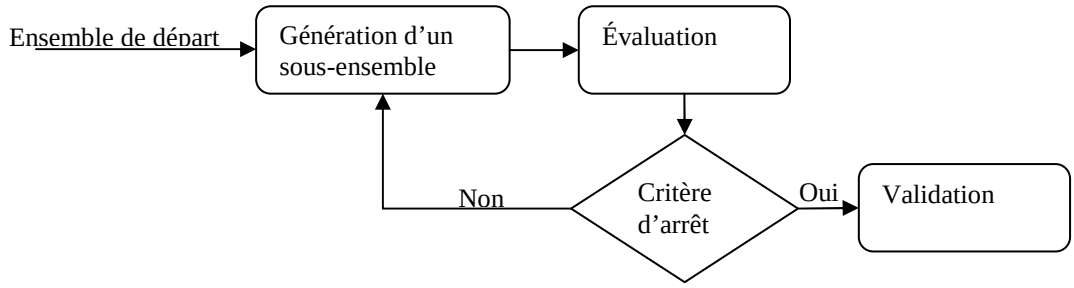


Figure 4.1. Procédure générale de sélection de caractéristiques

Les catégories des méthodes de sélection ou de pondération de caractéristiques les plus fréquentes dans la littérature, sont les méthodes de filtrage, les méthodes enveloppantes et les méthodes intégrées comme le synthétise la figure 2.

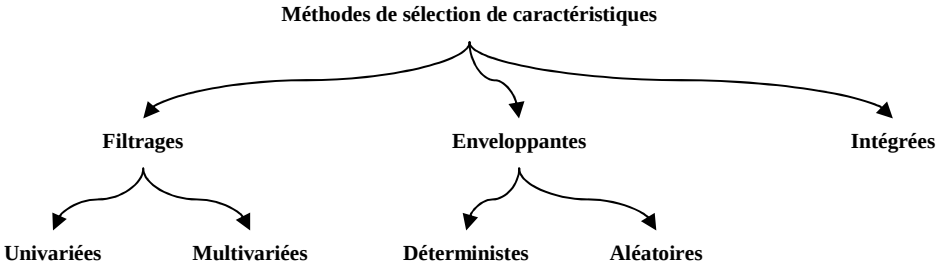


Figure 4.2. Les catégories de méthodes de sélection de caractéristiques

Les méthodes de Filtrage : Dans les méthodes de filtrage *(Dash et Liu, [1997])* la performance de chacune des caractéristiques est évaluée individuellement et un score de pertinence est attribué à chacune d'elles. Celles qui ont le score le plus faible sont supprimées. Le sous-ensemble de caractéristiques restant va constituer l'entrée de l'algorithme de classification. Les avantages des méthodes de filtrage sont qu'elles peuvent facilement s'adapter aux données de grande dimension : elles offrent des possibilités de calculs simples et rapides et sont indépendantes du classifieur *(Chapelle et Vapnick, [1999]).* En conséquence, la sélection de caractéristiques est appliquée une seule fois puis est évaluée à partir de différents classifieurs *(Saeys et al. [2007]).* La figure 4.3 illustre le modèle de sélection de caractéristiques par les méthodes filtres. Dans ce modèle la sélection de caractéristiques est considérée comme une étape de prétraitement.

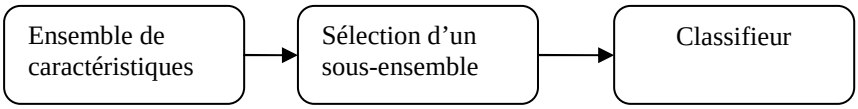


Figure 4.3. Modèle général de sélection de caractéristiques à partir des méthodes de filtrage. Les caractéristiques sont filtrées indépendamment de l'algorithme d'induction

Un inconvénient souvent relevé des méthodes de filtrage est qu'elles ignorent l'interaction avec le classificateur et la plupart des méthodes proposées sont *univariées*, cela signifie que chaque caractéristique est examinée séparément, ignorant ainsi la dépendance avec les autres caractéristiques, ce qui constitue un frein souvent majeur aux bonnes performances de classification comparativement à d'autres méthodes de sélection. Afin de résoudre le problème de l'indépendance des caractéristiques, des méthodes de filtrage *multivariées* ont été proposées visant ainsi à l'incorporation de dépendances entre caractéristiques.

Les méthodes enveloppantes « wrapper » :

Dans les méthodes enveloppantes, la sélection d'un sous-ensemble de caractéristiques est appliquée en utilisant un classifieur. L'algorithme de sélection d'un sous-ensemble de caractéristiques effectue une recherche d'un sous-ensemble de caractéristiques pertinentes en utilisant le classifieur comme une fonction d'évaluation de ce sous-ensemble. La précision du classifieur est estimée en utilisant des techniques d'évaluation comme bootstrap ou cross-validation *(Kohavi, [1995])* (figure 4.4).

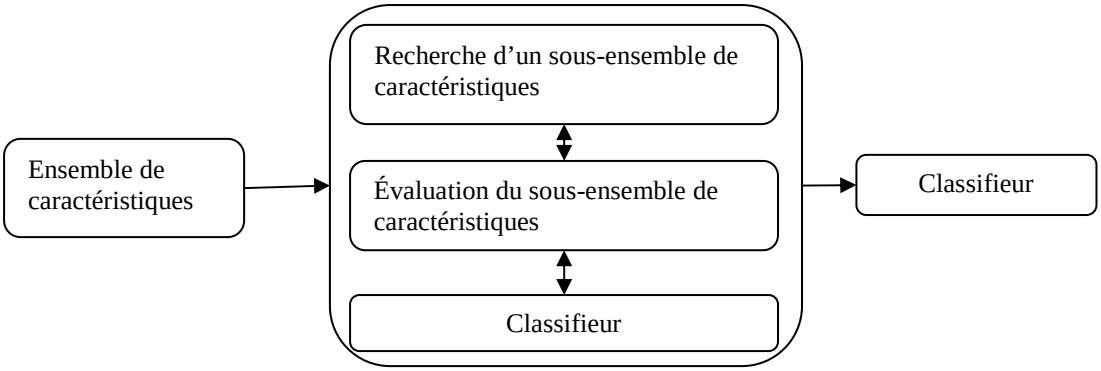


Figure 4.4. Modèle général de sélection de caractéristiques à partir des méthodes enveloppantes

Les méthodes enveloppantes demandent un grand temps de calcul à cause de la complexité de l'algorithme d'apprentissage. Pour diminuer le temps de calcul et éviter le sur-apprentissage, le processus de validation croisée *(Mitchel, [1996])* est souvent utilisé. La complexité de cette méthode ne permet pas l'utilisation des méthodes de recherche exhaustives (problème *NP*-complet). Pour cela les méthodes de recherche heuristique sont utilisées pour guider la recherche vers un sous-ensemble optimal. Dans *(Li et Guo, [2008]), (Huang et al. [2008])* les

auteurs ont montré que les méthodes enveloppantes produisaient de meilleurs résultats que les méthodes de filtrage. Les principaux inconvénients de ces méthodes enveloppantes sont liés d'une part à la complexité de calcul où le temps demandé pour chercher un sous-ensemble de caractéristiques optimales est plus élevé que dans le cas des méthodes de filtrage, et d'autre part aux spécificités du classifieur : les résultats de sélection peuvent changer d'un classifieur à un autre.

Les méthodes intégrées: Dans cette famille de méthodes, nous ne pouvons pas obtenir de séparation entre la partie apprentissage et la partie de sélection des caractéristiques, ce qui diffère des autres méthodes de sélection de caractéristiques. Comme les méthodes enveloppantes, ces méthodes sont spécifiques à un classifieur. L'avantage des méthodes intégrées est donc lié au fait qu'elles incluent l'interaction avec le modèle de classification, réduisant ainsi les temps de calcul par rapport aux méthodes enveloppantes *(Saeys et al. [2007])*. Le tableau 4.1 résume les trois types de méthodes de sélection de caractéristiques.

Tableau 4.1. Résumé des méthodes de sélection de caractéristiques

Approche	Avantages		Désavantages	Méthodes
Filtre	**Univariée**	-Rapide -Indépendant du classifieur -Évolutive	-Ignore la dépendance entre les caractéristiques -Ignore l'interaction avec le classifieur	-Chi-deux -Distance euclidienne -T-test
	Multivariée	-Modélise la dépendance des caractéristiques -Indépendant du classifieur -Meilleure complexité de calcul que les méthodes de wrapper	-Plus lent que les techniques univariés -Moins évolutive que les techniques univariés -Ignore l'interaction avec le classifieur	-Sélection de caractéristiques basée sur la correlation -Filtre Markov -Sélection de caractéristiques basée sur une corrélation rapide
Enveloppante	**Déterministe**	-Simple -Interagit avec le classifieur -Modélise la dépendance entre les caractéristiques -Moins gourmande en ressources que les méthodes radomisés	-Sélection dépendante du classifieur -Plus enclins à être coincé dans un optimum local que les algorithmes randomisés	-Sequential forward selection (SFS) -Sequential backward Selection (SBS) -Beam Search
	randomisé	-Intéragit avec les classificateurs -Modélise la dépendance entre les caractéristiques	-Calcul intensif -Sélection dépendant du classifieur -Un risque plus grand de sur apprentissage que les algorithmes déterministes	-Simulated annealing -Algorithmes génétiques
Intégrée	-Intéragit avec le classifieur -Meilleur complexité de calcul que les méthodes de wrappers -Modélise la dépendance de caractéristiques		Sélection dépendante du classifieur	-Arbres de décisions -Réseaux bayésiens -Sélection de caractéristiques en utilisant les Machines à support Vecteur

2.2 Sélection et pondération des caractéristiques

La sélection de caractéristiques sélectionne un sous ensemble de caractéristiques qui permettront d'améliorer la performance de la classification, tandis que les caractéristiques sélectionnées gardent leur interprétation physique original *(Jain et al. [2000]).* Ces problèmes sont considérés comme *NP-difficile (Amaldi et Kann, [1998]).* La sélection et la pondération de caractéristiques sont implémentées comme la recherche de la solution optimale approximative *(Raymer et al. [2000]), (Cover et Campenhout, [1997]).*

La pondération de caractéristiques est suffisante pour résoudre des problèmes où les caractéristiques varient en pertinence. La sélection de caractéristiques donne de bons résultats lorsque les caractéristiques qui décrivent les instances sont soit fortement corrélées au contenu de la classe, soit non pertinentes mais en grand nombre *(Wettschereck et al. [1997]).* Dans ce contexte, de nombreux algorithmes ont été proposés pour trouver des solutions optimales approximatives. Le problème de sélection de caractéristiques est lié à la taille de l'ensemble de caractéristiques : un petit ensemble de caractéristiques contient environ *20* caractéristiques, un ensemble moyen en contient environ de *20* à *49* caractéristiques et un grand ensemble peut en contenir plus de *50 (Zhang et Sun, [2002]).*

Les algorithmes Sequential Forward Floation Search (SFFS) et Sequential Backward Floation Search (SBFS) donnent de bons résultats quand ils sont appliqués sur des ensembles de caractéristiques de petite ou grande taille *(Pudil et al. [1995]),* mais ils sont confrontés au problème de convergence vers des optimums locaux dès qu'ils sont utilisés sur de grands ensembles de caractéristiques *(Zhang et Sun, [2002]).*

Les algorithmes itératifs heuristiques comme les AG offrent de bonnes performances quand ils sont appliqués à la résolution de problèmes qui ont un espace de recherche de très grande dimension et bruité avec de nombreux optimums locaux caractéristiques *(Zhang et Sun, [2002]), (Sait, et Youssef, [2002]).* Les AG sont considérés comme des algorithmes évolutifs quand ils sont utilisés pour résoudre le problème de sélection de caractéristiques. Ils sont utilisés pour améliorer la performance de la classification *(Chouaib et al. [2009]), (Rafat et Soryani, [2006]),* mais ne garantissent pas de fournir une solution exacte optimale, et en général seule une solution quasi-optimale est fournie avec un temps de calcul raisonnable *(Sivaraj et al. [2012]).*

2.4 Principe général des algorithmes génétiques utilisés pour la sélection de caractéristiques

Les algorithmes génétiques (GA) constituent une méthodologie générale de recherche adaptative d'optimisation basée sur une analogie directe avec la sélection naturelle darwinienne et la génétique dans les systèmes biologiques. Il a été prouvé qu'ils constituaient une alternative prometteuse aux méthodes heuristiques classiques. Basé sur le principe darwinien de «survie du plus apte», l'AG travaille avec un ensemble de solutions candidates appelé *population* et obtient une solution optimale après une série de calculs itératifs.

L'AG évalue la fitness de chaque individu c.à.d. la qualité de la solution, grâce à une fonction de fitness. Les chromosomes qui ont de meilleures fitness ont plus de probabilité d'être conservés dans la prochaine génération. Si le meilleur individu ou chromosome dans une population ne peut pas satisfaire les exigences, des populations successives seront produites pour fournir des solutions alternatives. Les fonctions de croisement et de mutation sont les principaux opérateurs qui transforment d'une manière aléatoire les chromosomes et finalement impactent leur valeur de fitness. La population évoluera jusqu'à ce que des résultats acceptables soient obtenus. L'algorithme génétique peut être adapté efficacement à de grands espaces de recherche, et a donc moins de chance de s'arrêter sur une solution optimale locale que d'autres algorithmes.

La figure 4.5 illustre les opérateurs de croisement et de mutation classiques. L'opérateur de croisement recompose les gènes d'individus existant dans la population, il mélange les chromosomes des individus parents pour créer le code génétique d'un individu enfant. En mutation des gènes peuvent parfois être modifiés, par exemple, en changeant la valeur du gène de *0* à *1* ou vice versa dans un chromosome binaire *(Goldberg, [1989]), (Davis, [1991]).*

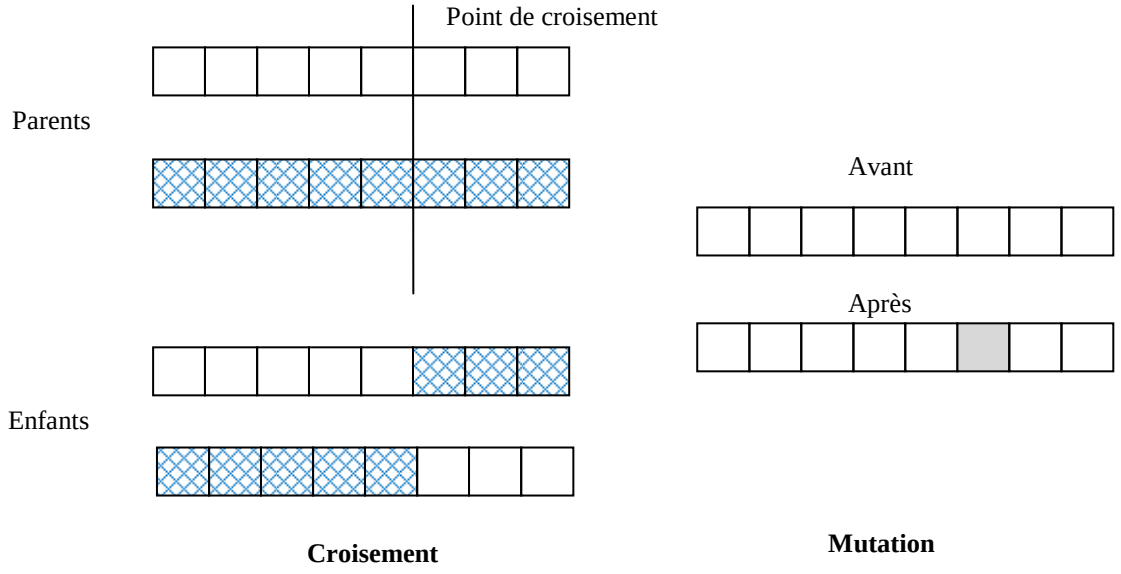


Figure 4.5. Illustration des opérateurs de croisement et mutation *(Huang et Wang, [2006])*

L'algorithme suivant décrit une version générique d'un algorithme génétique *(Russell et Norvig, [2003])* :

Algorithme Génétique

1. Soit $s_1,\ldots s_N$ la population actuelle
2. Soit $p_i = f(s_i) \Big/ \sum_j f(s_j)$ la probabilité qu'un chromosome soit sélectionné pour la reproduction.
3. Pour $k = 1 ; k < N ; k += 2$
   a. *Parent 1* : choisir au hasard en fonction de *p*
   b. *Parent 2 :* choisir au hasard en fonction de *p*
   c. choisir au hasard un point de croisement, échanger des chaînes de parents *1, 2* pour générer les enfants *t[k], t [k +1]*
4. Pour $k = 1 ; k <= N ; k++$
   a. Aléatoire muter chaque position en *t [k]* d'une faible probabilité (taux de mutation)
5. La nouvelle génération remplace l'ancien: $\{s\} \leftarrow \{t\}$
6. Répéter.

3 Construction d’un dictionnaire de formes associé à un manuscrit

3.1 Principe général de l’approche de construction de dictionnaire de formes par AG

Les différentes étapes de notre démarche de clustering de graphèmes sont présentées par le synoptique la (figure 6(a)). Chaque graphème est représenté par un ensemble de caractéristiques. On utilise l’AG pour définir une sélection automatique du meilleur sous-ensemble de caractéristiques des graphèmes associés au meilleur seuil de classification. Ce seuil contrôle le calcul des similarités entre un graphème et un cluster représenté par son graphème barycentre (la moyenne des graphèmes du cluster). Pendant l'évolution de l'optimisation (mesurée par une fonction de fitness détaillée plus tard), le module AG coopère en permanence avec le module de classification de la façon suivante :

- En entrée, l’algorithme génétique envoie au classifieur différentes générations de paramètres (le sous-ensemble de caractéristiques et le seuil de classification par coloration).
- A chaque génération de paramètres, le classifieur classe les graphèmes en groupes induisant un dictionnaire de formes.
- La qualité du résultat de chaque classification est évaluée par une fonction de fitness.
- Cette fonction fournit à l’AG des indications sur les sélections efficaces et permet ainsi au système de proposer une nouvelle pondération de paramètre qui sera acheminée à nouveau vers l’entrée du classifieur.
- Le processus d’échanges mutuels entre les deux modules (AG/Classifieur) est répété jusqu'à la stabilisation ou jusqu'à ce qu’un nombre d’itérations maximal (défini par l’utilisateur) soit atteint.

Les différentes étapes de génération de poids générique des caractéristiques sélectionnées sont représentées dans la figure 4.6 (b). Les poids génériques et un seuil sont déduits à partir d’une base d’apprentissage (section 3.7). Ceci produit un ensemble de poids génériques et un seuil optimal qui permettra de résoudre les contraintes suivantes dans notre travail

- Le problème des rejets des caractéristiques dans le cas de la sélection binaire par AG est considéré comme sévère sur les caractéristiques qui ont un faible pouvoir discriminant. Au lieu de cela, elles seront affectées de faibles valeurs de poids et la

caractéristique avec un pouvoir discriminant élevé sera affectée des valeurs de poids plus élevées.

- Cela va produire un vecteur générique de poids *w**. Ainsi, nous aurons seulement à appliquer le vecteur générique de poids aux caractéristiques représentant un nouveau manuscrit que nous voulons entrer dans notre étude comparative.
- Il nous permettra également d'avoir un seuil d'adjacence qui est la moyenne de tous les seuils qui sont donnés pour chaque manuscrit dans notre base d'apprentissage.

Afin de prouver que notre proposition est efficace, c.à.d. que les fonctions de fitness qui résultent des poids génétiques sont proches des fonctions de fitness de la sélection ou même plus élevées, nous avons réalisé quelques expériences en utilisant la base de manuscrits médiévaux d'Oxford.

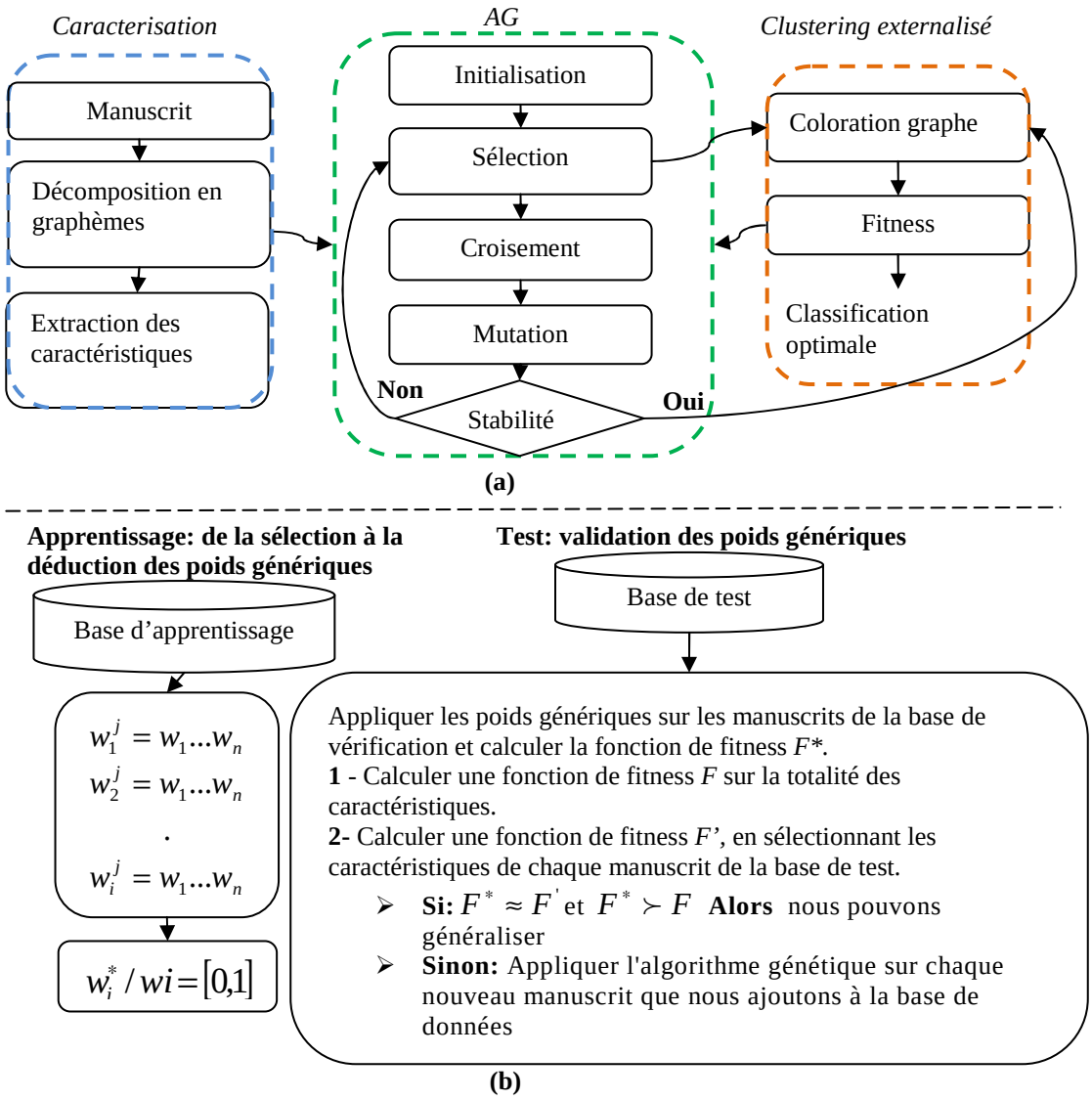


Figure 4.6. (a) Caractérisation et **sélection de caractéristiques par algorithme génétique combiné à la coloration de graphes, (b) déduction des poids et test de validité du système**

3.2 Choix des descripteurs initiaux

Initialement, nous produisons une description vectorielle de chacun des graphèmes qui servira au module de clustering définissant l'ensemble des entrées du dictionnaire de formes. Les *59* caractéristiques utilisées pour la description des graphèmes sont groupées en *12* descripteurs de la manière suivante : {D_1=*Hauteur*, D_2=*Largeur*, D_3=*Excentricité*, D_4=*Densité Globale*, D_5= *Direction*, D_6=*Périmètre*, D_7=*circularité*, D_8=*Compacité*, D_9 = *9 Densités*, D_{10}= *Huit orientations*, D_{11}=*direction des 9 blocs*, D_{12} = *Moments de Zernike*}. Ces descripteurs ont été choisis parce qu'ils quantifient des informations visuelles fréquentes de forme des graphèmes (informations géométrique, directionnelles, topologiques). Ce vecteur de

caractéristiques est le résultat de la concaténation des caractéristiques géométriques et des moments de Zernike.

3.2.1 Les descripteurs et caractéristiques choisis

Caractéristiques géométriques

- **Les descripteurs scalaires (figure 4.7) :**
  - La hauteur D_1 et la largeur D_2 sont utilisées pour quantifier l'élongation du trait
  - L'excentricité D_3 représente le ratio entre la hauteur et la largeur.
  - La densité globale D_4 est utilisée pour savoir le nombre de pixels constituant le graphème, cette mesure est autrement un indicateur de l'épaisseur de la tête du calame utilisé pour la formation des traits.
  - La direction D_5 permet de connaître l'inclinaison du graphème et de différencier les mouvements d'exécution de l'écriture.
  - Le périmètre D_6 décrit la longueur de la ligne qui délimite le contour d'un objet.
  - La circularité D_7 décrit la forme d'un objet selon la formule suivante:

$$4 \times \pi \times Aire \Big/ périmètre^2 \qquad \textbf{(4.1)}$$

  - La compacité D_8 décrit la forme d'un objet selon la formule suivante :

$$C = \sqrt{(4/\pi) \times Aire} \Big/ Hauteur \qquad \textbf{(4.2)}$$

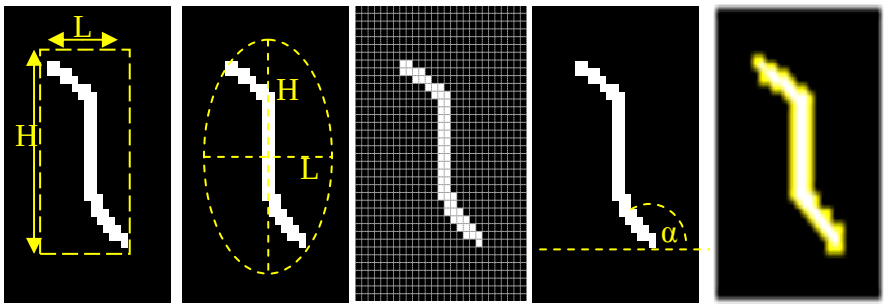


Figure 4.7. (De gauche à droite), Hauteur et largeur d'un graphème, excentricité, densité globale, direction, périmètre

- **Les descripteurs représentés par des vecteurs de dimension *n***
    - Les 9 densités D_9. Le même principe que le descripteur D_4, mais cette fois l'image est divisée en 9 régions et nous calculons l'aire pour chacune des régions.
    - D_{10}, est un vecteur de 8 compostantes chacune associée à une direction $\theta_i(0°,45°,90°,135°,180°,225°,270°,315°)$. On considère les segments issus de chaque pixel contour, chacun contribue à D_{10} en fonction de sa direction. Chaque composante est normalisée par le nombre total de pixels du contour.
    - Les directions des *9* blocs (zones) D11 représentent la forme des courbes de graphèmes et reflètent les propriétés structurelles telles que la convexité et de concavité. Dans chacun des *9* blocs, les points avec des extrema locaux de la courbure sont considérés comme candidats dominants.

Moments de Zernike

Les Moments de Zernike ont été introduits par *(Teague, [1980]).* Ils sont construits à partir de polynômes complexes et forment un ensemble orthogonal complet défini sur le disque unité. Ces moments sont invariants à la rotation et robustes aux bruits. Les *25* moments de Zernike D_{12} utilisés pour décrire les graphèmes sont classés parmi les moments orthogonaux (géométrique, de Legendre, etc.) car ils possèdent la propriété d'invariance à la rotation. Le moment de Zernike d'ordre *n* avec la répétition m (n - |m| est paire et |m|<n) est défini par:

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x,y) \cdot V^*_{nm}(\rho,\theta) \qquad \textbf{(4.3)}$$

$V_{nm}(\rho,\theta)$ est un ensemble de polynômes complexes dans l'espace à deux dimensions qui forme un ensemble orthogonal sur l'intérieur du disque unité ouvert $(x^2+y^2=1)$, avec :

$$V_{nm}(\rho,\theta) = R_{nm}\ e^{im\theta} \qquad \textbf{(4.4)}$$

où ρ est la longueur du vecteur d'origine au point de coordonnées (x , y). θ est l'angle entre le précédent vecteur et l'axe des abscisses. $R_{n,m} = R_{n,-m}$ est un polynôme radial défini comme suit :

$$R_{nm}(\rho) = \sum_{S=0}^{(n-|m|)/2} (-1)^S \frac{(n-S).!}{S!\left(\frac{n+|m|}{2}-S\right)!\ \left(\frac{n-|m|}{2}-S\right)!} \rho^{n-2S} \qquad \textbf{(4.5)}$$

Pour calculer les moments de Zernike, le centre de chaque graphème binaire est pris comme origine du repère et les coordonnées des pixels de l'image sont transformées de manière à être dans le domaine du cercle unité. Comme nous l'avons signalé précédemment les moments de Zernike sont invariants seulement à la rotation. Pour les rendre invariants au changement d'échelle *(Teague, [1980]),* il est nécessaire de normaliser l'image binaire du graphème par le moment du premier ordre m_{00} défini comme étant l'aire du graphème.

Cette propriété d'invariance à la rotation est généralement importante en analyse d'images, cependant elle ne sera pas directement exploitée dans nos travaux car du point de vue de l'exécution des traits, nous cherchons précisément à différencier les formes concaves des formes convexes qui traduisent chacune des mouvements d'écriture différents. L'invariance à l'échelle en revanche est un critère important.

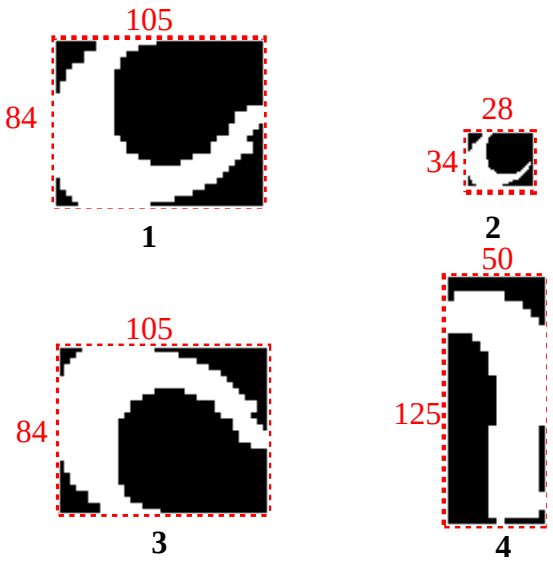
3.2.1.1 Evolution des valeurs des descripteurs en fonction des graphèmes

Dans ce paragraphe nous montrons l'évolution des valeurs des descripteurs pour chacune des trois familles: scalaires, vecteurs de *n* dimensions et moments de Zernike, en représentant quel effet peuvent produire les changements de formes de graphèmes sur les valeurs de ces derniers. Nous terminerons par un bilan sur ces mesures montrant leur invariance possible à la rotation et à l'échelle. Dans cette étude nous nous intéressons par une description complète (rondeur, linéarité, dimension...) de formes indépendamment de leurs propriétés d'invariance à l'échelle et à la rotation, car ces éléments ne sont pas des facteurs essentiels de discrimination entre graphèmes. Néanmois nous présentons ces propriétés pour montrer qu'elles pourraient être exploitées sur d'autres types de contenus, ou prises individuellement en fonction de leurs propriétées, par exemple l'invariance à l'échelle et à la rotation sont des propriétés qui pourraient être exploitées lors d'une sélection manuelle des descripteurs.

Les descripteurs scalaires

La hauteur D_1 et la largeur D_2

D_1 appartient à l'intervalle *[1, H]* où *H* est la hauteur de l'imagette du graphème. D_2 est comprise dans l'intervalle *[1, L]* où *L* est la largeur de l'imagette (figure 4.8).



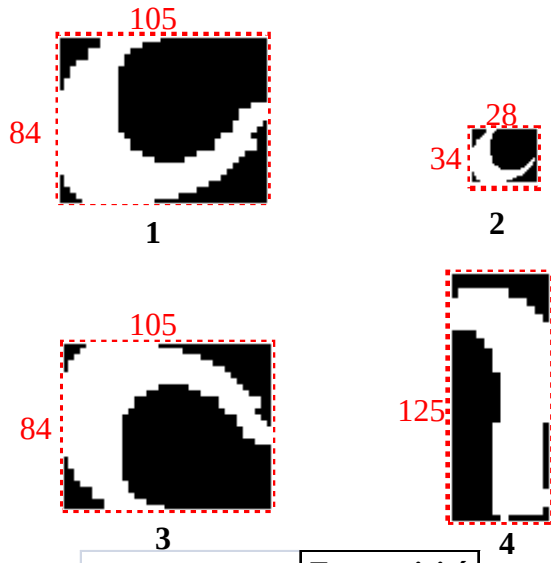
	Hauteur	Largeur
Graphème 1	84	105
Graphème 2	28	34
Graphème 3	84	105
Graphème 4	125	50

Figure 4.8. Evolution des graphèmes en fonction de la hauteur et de la largeur

Les exemples de la figure 4.8 montrent que la hauteur et la largeur ne sont pas dépendantes de la forme du graphème. Elles nous fournissent seulement comme nous avons défini avant une information sur l'élongation des graphèmes. Notons que les dimensions prises individuellement de hauteur et de largeur ne conduisent pas à une invariance en échelle comme le montre les graphèmes *1* et *2*. Mais peuvent conduire à une invariance en rotation (deux graphèmes orientés différemment peuvent avoir les deux mêmes dimensions de hauteur et de largeur, comme le montre les graphèmes *1* et *3*).

L'Excentricité D_3

L'excentricité est égale au rapport $D_1(Hauteur)/D_2(L\arg eur)$ (figure 4.9).



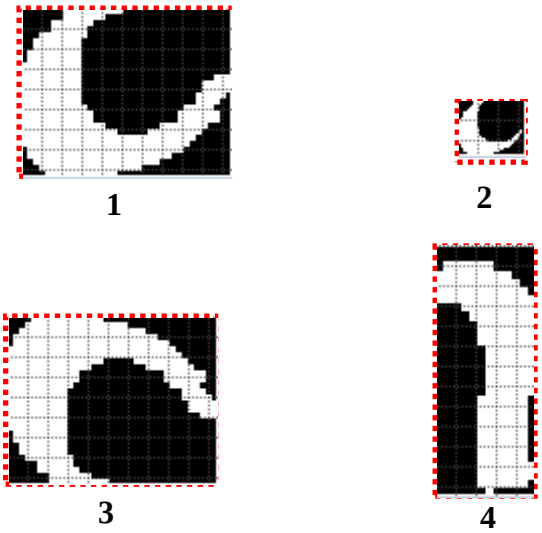
	Excentricité
Graphème 1	1,25
Graphème 2	1,21
Graphème 3	1,25
Graphème 4	0,4

Figure 4.9. Evolution de l'excentricité des graphèmes en fonction de transformations

La figure 4.9 montre les valeurs de l'excentricité qui résultent du rapport Hauteur/Largeur. Ces descripteurs simples peuvent seulement discriminer des formes avec de grandes différences, et ne peuvent pas etre utilisés comme les seules descripteurs pour appliquer une discrimination entre les formes. Notons aussi aussi que l'excentricité est invariante à la rotation comme le montre les graphèmes *1* et *3*. Mais elle n'est pas invariante à l'échelle, après la réduction de la hauteur et de la largeur de 1er graphème nous remarquons que la valeur de l'excentricité change de *1,25* à *1,21*.

Densité globale D_4

La densité globale représente l'aire du graphème et elle est comprise entre *[5, n]* où *n* est le nombre de pixels (blancs) qui constituent le graphème (figure 4.10).



	Densité globale
Graphème 1	4055
Graphème 2	429
Graphème 3	4055
Graphème 4	3584

Figure 4.10. Evolution de la densité globale des graphèmes en fonction de transformations

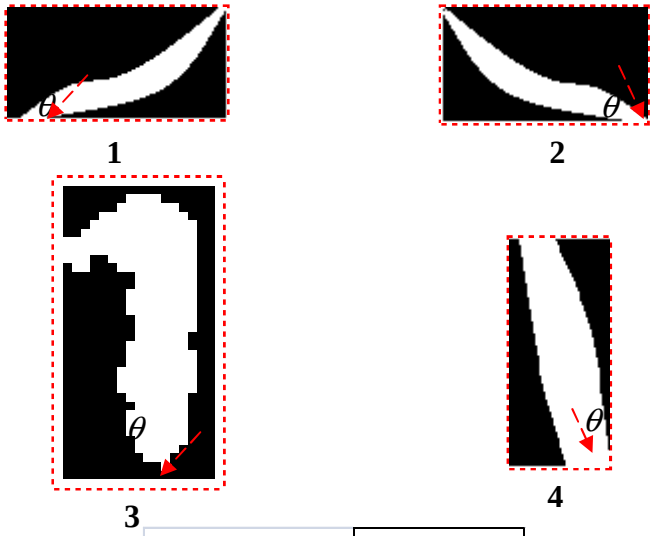
La densité globale fournit une indication d'épaisseur du graphème et elle donne une indication sur le style de l'écriture, comme cela s'observe pour les écritures gothiques et carolingiennes. Ces dernières ont une épaisseur de trait moins importante que les premières.

La figure 4.10 montre que la densité globale est invariante à la rotation mais pas à l'échelle. Les graphèmes *1* et *3* ont la même densité globale après avoir appliqué une rotation de *180°*. Mais quand il y a un changement de taille, la densité globale de la forme change. La densité ainsi définie n'est pas contrôlée par les dimensions de l'imagette, c'est pour cette raison qu'elle n'est pas invariante à l'échelle. Deux graphèmes de même surface auront donc la même valeur de densité D_4. Ce descripteur pris indépendamment des autres ne relève donc ni des différences morphologiques, ni des variations d'orientations. Il permet cependant de proposer un ordonnacement des graphèmes selon un critère de surface.

.

La Direction D_5

La direction d'un graphème est comprise entre *[0°,360°]* (figure 4.11).



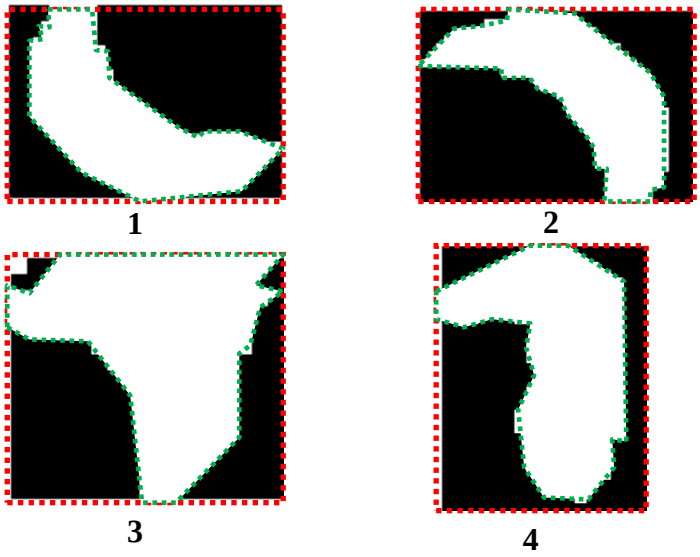
	Direction
Graphème 1	358,99
Graphème 2	1
Graphème 3	9,94
Graphème 4	16,78

Figure 4.11. Evolution de la direction des graphèmes en fonction de transformations

La figure 4.11 montre que la direction donne une indiquation de l'inclinaison du graphème. Elle nous permet de distinguer les différents styles d'écritures. L'écriture de style carolingien est plus inclinée que celle de style gothique. Notons la direction est invariante à l'échelle mais pas à la rotation, en effet pour un graphème nous avons toujours une seule valeur pour la direction même après le changement de taille, mais après une rotation la direction évolue en fonction de la valeur de l'angle de rotation θ.

Le Périmètre D_6

Le périmètre du graphème dépend de la taille et il est compris entre *[5, n]*, où *n* présente le nombre de pixels du contour (figure 4.12).



	Périmètre
Graphème 1	217
Graphème 2	217
Graphème 3	369
Graphème 4	302

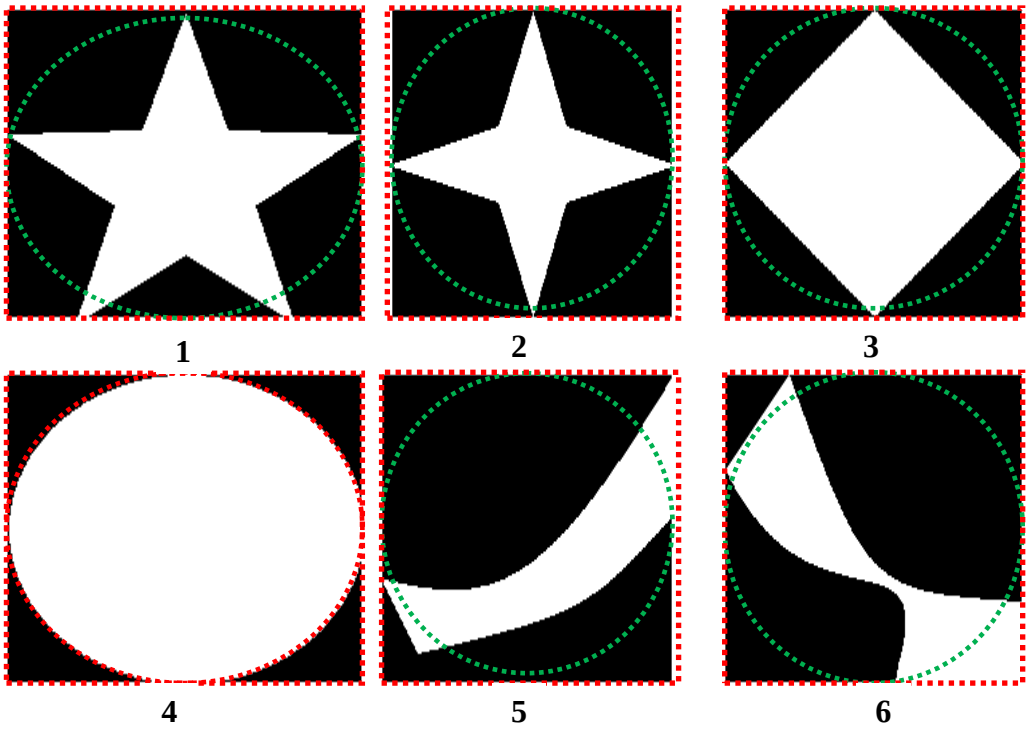
Figure 4.12. Evolution du périmètre des graphèmes en fonction de transformations

La figure 4.12 montre que le périmètre est redondant avec les dimensions de hauteur et de largeur : nous verrons que son impact pour produire un clustering optimal des graphèmes en classes de formes sera très faible. Nous avons néanmoins conservé cette dimension géométrique simple car elle fait partie d'un ensemble standard de descripteur géométrique de petites formes.

La figure 4.12 montre aussi que le périmète est invariant à la rotation mais pas au changement d'échelle, en effet le premier et second graphème ont le même perimètre après la rotation du graphème.

La Circularité D_7 et La compacité D_8

La circularité et la compacité sont comprises entre *[0, n]* (figure 4.13).



	Circularité	Compacité
Graphème 1	0,20	0,56
Graphème 2	0,21	0,48
Graphème 3	0,40	0,68
Graphème 4	0,82	0,84
Graphème 5	0,26	0,48
Graphème 6	0,31	0,50

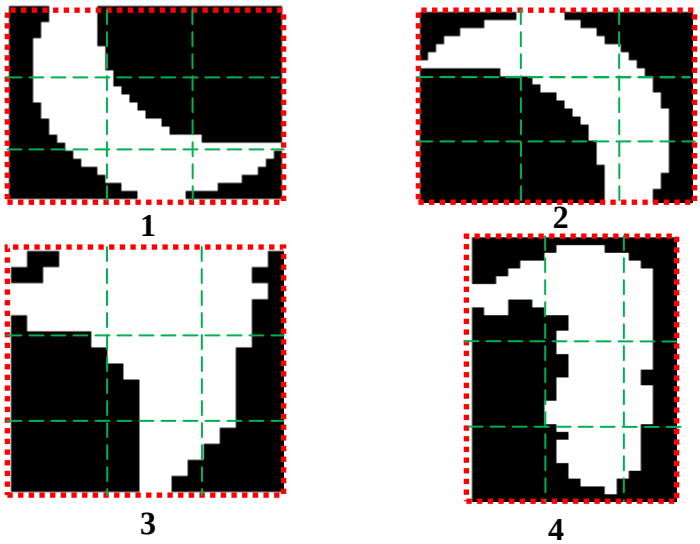
Figure 4.13. Evolution de la circularité et de la compacité des graphèmes en fonction de transformations

La figure 4.13 illustre l'évolution de formes (prises volontairement différentes des graphèmes pour bien visualiser l'échelle des diversités possibles) en fonction de la circularité et compacité. Les valeurs de la circularité montrent que plus la forme devient circulaire, plus elle devient proche de *1* et plus la forme devient longitudinale plus elle décroit et devient proche de *0*. La compacité suit le même principe, plus la forme est compacte plus elle est proche *1* et plus elle s'allonge plus elle décroit. Ces deux valeurs nous aident à différencier les styles d'écritures puisque l'écriture carolingienne comprent plus de formes allongées que le style gothique où les formes sont plus compactes. La circularité et la compacité sont aussi invariantes à la rotation mais pas à l'échelle.

Les descripteurs représentés par un vecteur de *n* dimension

Les 9 densités D_9

Les *9* densités forment un vecteur $V=(n_1,...n_9)$ où n_i présente l'air de chacune des *9* zones de l'image du graphème (figure 4.14).



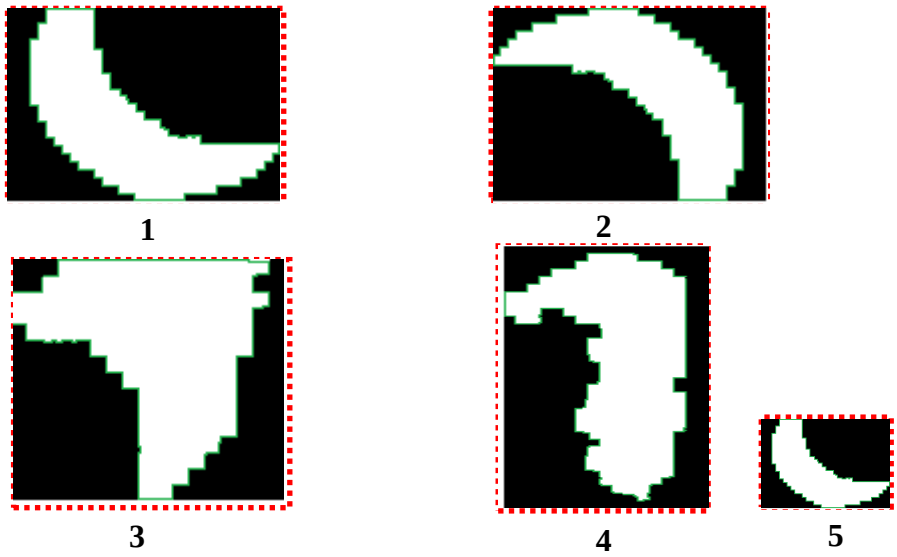
	1	2	3	4	5	6	7	8	9
Graphème 1	0	25	5	0	25	5	5	25	9
Graphème 2	9	70	48	50	30	7	52	0	7
Graphème 3	50	0	7	52	23	7	14	75	48
Graphème 4	28	44	40	0	23	45	9	42	7

Figure 4.14. Evolution des 9 densités des graphèmes en fonction de transformations

Comme la densité globale, les *9* densités nous donnent une indication sur l'épaisseur des traits des différents styles d'écriture. Ce descripteur va apporter des informations importantes sur l'occupation du graphème dans la surface de sa boîte englobante (figure 4.14). Ces 9 densités ne sont pas invariantes à la rotation comme le montre les graphèmes *1* et *2*, ni à l'échelle puisque le calcul de la densité consiste au dénombrement de pixels objets sur *9* zones.

Les *8* orientations D_{10}

Les *8* orientations forment un vecteur $V=(\theta_1,...\theta_8)$ où θ_i présentent le nombre de segments dans l'une des *8* directions (figure 4.15).



	0°	*45°*	*90°*	*135°*	*180°*	*225°*	*270°*	*315°*
Graphème 1	40	45	14	48	49	14	45	41
Graphème 2	14	45	41	48	49	40	45	14
Graphème 3	12	26	15	18	16	16	26	11
Graphème 4	30	64	30	39	37	30	64	29
Graphème 5	27	34	8	39	39	9	34	28

(a)

	0°	*45°*	*90°*	*135°*	*180°*	*225°*	*270°*	*315°*
Graphème 1	0,57	0,66	0,11	0,71	0,73	0,11	0,66	0,59
Graphème 2	0,11	0,66	0,59	0,71	0,73	0,57	0,66	0,11
Graphème 3	0,07	0,32	0,13	0,18	0,14	0,14	0,32	0,05
Graphème 4	0,39	1,00	0,39	0,55	0,52	0,39	1,00	0,38
Graphème 5	0,34	0,46	0,00	0,55	0,55	0,02	0,46	0,36

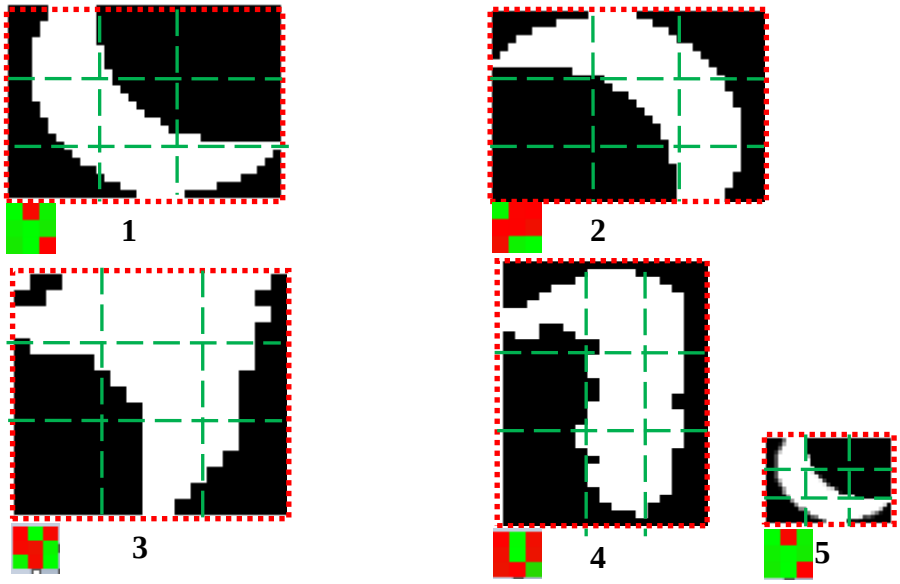
(b)

Figure 4.15. Calcul de D_{10} pour *5* graphèmes, les résultats sont donnés (a) avant normalisation, (b) après normalisation

D'après le tableau, nous remarquons que les graphèmes *1*,*2* et *5* présentent de plus grandes similitudes de valeurs par rapport aux graphèmes *3* et *4*. Les expérimentations montrent que ces caractéristiques sont très pertinentes pour la discrimination des styles d'écritures paléographiques (figure 4.15 (b)). Nous remarquons également d'après les valeurs du tableau que ces orientations sont invariantes à l'échelle mais pas à la rotation.

Les directions des *9* blocs D_{11}

Les *9* directions forment un vecteur $V = (\alpha_1, ...\alpha_{n=9})$ où les valeurs de α_i présent les *9* directions dominantes du contour (figure 4.15).



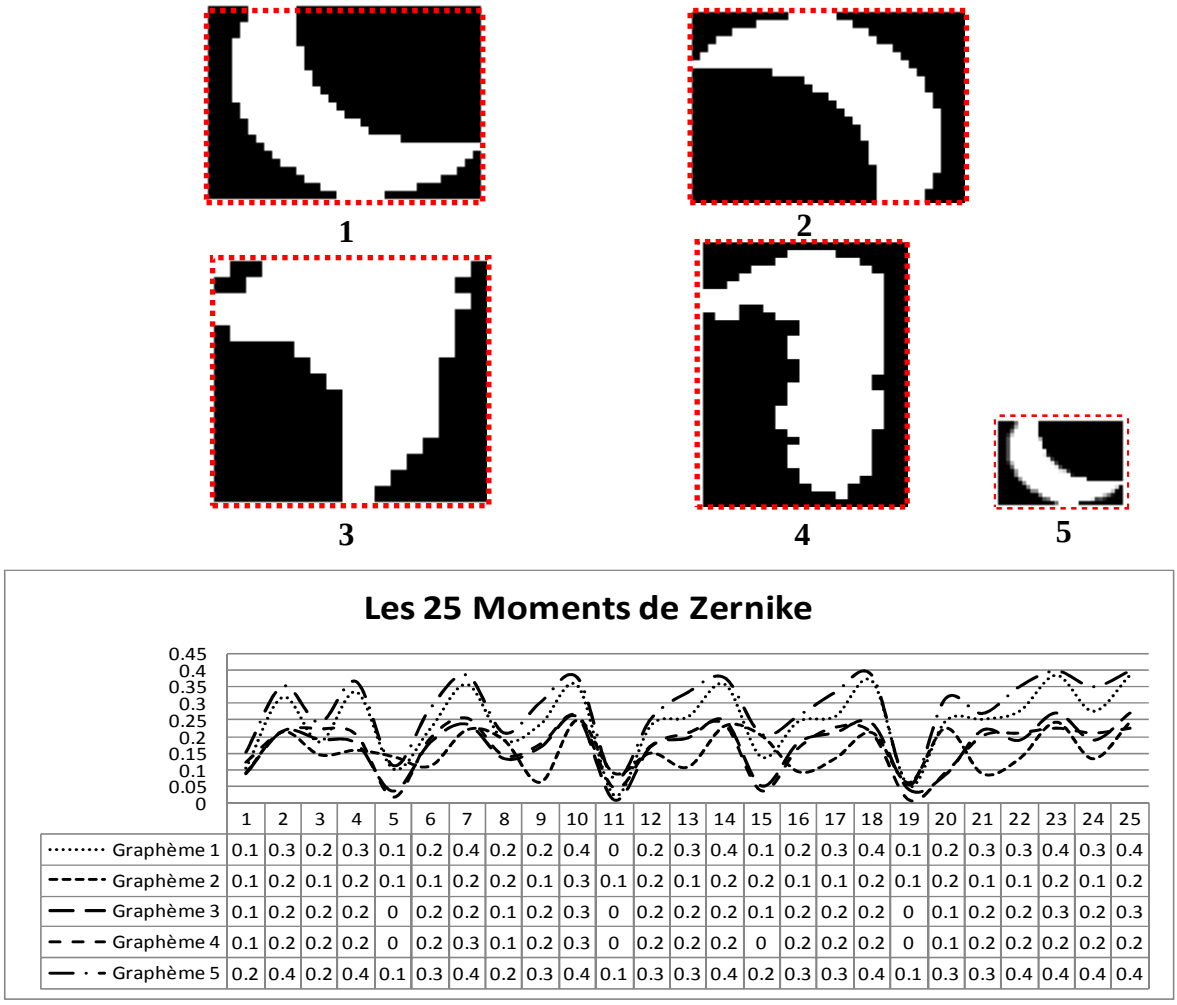
Graphème 1	2	343	19	2	30	0	351	6	354
Graphème 2	9	356	360	357	339	18	2	304	0
Graphème 3	360	0	349	334	16	343	0	0	333
Graphème 4	356	4	337	0	331	357	54	345	35
Graphème 5	2	340	20	3	27	0	351	3	2

Figure 4.16. Evolution des graphèmes en fonction des *9* orientations, avec les cartes de chaleurs représentant les *9* directions pour chacun des 5 graphèmes

La figure 4.16, montre les *9* directions principles du contour des *9* zones de l'image. Ces *9* directions nous fourissent des informations sur la concavité et convexité des graphèmes. Nous remarquons aussi à partir des cartes de chaleurs que les *9* directions sont invariantes à l'échelle, comme le montre les graphèmes *1* et *5*, mais pas à la rotation comme c'est le cas des graphèmes *1* et *2*.

Les 25 Moments de Zernike

Les *25* Moments de Zernike forment un vecteur $M = (z_1,...,z_{n=25})$ où les valeurs de z_i présentent les *25* Moments.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Graphème 1	0.1	0.3	0.2	0.3	0.1	0.2	0.4	0.2	0.2	0.4	0	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.3	0.4	0.3	0.4
Graphème 2	0.1	0.2	0.1	0.2	0.1	0.1	0.2	0.2	0.1	0.3	0.1	0.2	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.2	0.1	0.2
Graphème 3	0.1	0.2	0.2	0.2	0	0.2	0.2	0.1	0.2	0.3	0	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0	0.1	0.2	0.2	0.3	0.2	0.3
Graphème 4	0.1	0.2	0.2	0.2	0	0.2	0.3	0.1	0.2	0.3	0	0.2	0.2	0.2	0	0.2	0.2	0.2	0	0.1	0.2	0.2	0.2	0.2	0.2
Graphème 5	0.2	0.4	0.2	0.4	0.1	0.3	0.4	0.2	0.3	0.4	0.1	0.3	0.3	0.4	0.2	0.3	0.3	0.4	0.1	0.3	0.3	0.4	0.4	0.4	0.4

Figure 4.17. Evolution des graphèmes en fonction des *25* Moments de Zernike

La figure 4.17 montre les *25* Moments de Zernike sur les *5* graphèmes. Nous remarquons que les valeurs sont très proches quand il existe un changement d'échelle comme dans le cas des graphèmes *1* et *5*, mais quand il y une rotation comme dans le cas des graphèmes *1* et *2*, nous constatons que les valeurs ne sont pas proches. Ce qui montre en effet l'invariance des Moments de Zernike par rapport à la rotation.

3.2.1.2 Bilan sur les 12 descripteurs

Dans le tableau suivant (tableau 4.3), nous présentons les descripteurs et précisons pour chacun d'eux s'il est invariant *(+)* ou dépendant *(-)* au changement d'échelle et à la rotation..

Tableau 4.3. Invariance des descripteurs à l'échelle et rotation

Descripteur	***Echelle***	***Rotation***
Hauteur	-	+
Largeur	-	+
Excentricité	+	+
Densité Globale	-	+
Direction	+	-
Périmètre	-	+
Circularité	-	+
Compacité	-	+
9 densités	-	-
8 orientations	+	-
Direction des *9* blocs	+	-
Moments de Zernike	+	-

Cela va produire :

- 2^n combinaisons pour la sélection binaire des caractéristiques ou des descripteurs
- 10^n combinaisons pour les pondérations (les pondérations appartiennent à l'intervalle *[0.1, 0.9])* avec des nombres réels de caractéristiques ou descripteurs, avec $n = 12$ pour les descripteurs et $n = 59$ pour les caractéristiques.

3.2.2 Le seuil d'adjacence

La classification est réalisée en utilisant l'algorithme de coloration de graphe. Cet algorithme utilise un seuil défini dans l'intervalle *[3, 90]*. L'intervalle des seuils a été choisi empiriquement quand nous avons commencé la construction des dictionnaires de formes en utilisant l'algorithme de coloration de graphe. Nous avons constaté que des valeurs de seuil inférieures à *3* conduisaient à une sur-segmentation des classes (à l'extrême une classe par graphème), et symétriquement des valeurs supérieures à *90* conduisaient à une sous-segmentation (et donc un nombre de classes insuffisant). Le seuil est ensuite converti en binaire pour être finalement intégré dans le codage d'un chromosome, ce qui va produire 2^7 valeurs possibles de seuil. Nous notons que la représentation binaire de la valeur seuil maximum de *90* est *1011010* (*7* bits).

3.3 Sélection ou pondération de caractéristiques par AG

Dans notre étude, nous avons été confronté à deux problèmes :

1. L'ajustement du seuil pour la coloration de graphe.
2. La sélection non-supervisée d'un sous-ensemble de caractéristiques calculées sur les graphèmes, offrant la meilleure représentation du dictionnaire de formes des graphèmes. Le seuil de coloration va influencer le nombre de classes produites par le classifieur alors que la sélection des caractéristiques va identifier les variables discriminantes qui nous permettrons de réaliser la meilleure séparation des graphèmes à classer. Nous présentons dans les paragraphes suivants les étapes de ces deux problèmes d'optimisation par l'algorithme génétique.

3.3.1 Représentation des Chromosomes

La première étape consiste à modéliser les variables entrant dans l'optimisation et le clustering, à savoir les *59* caractéristiques ou *12* descripteurs de chaque graphème ainsi que le seuil de coloration qui est codé sur *7* bits.

3.3.1.1 Représentation du chromosome pour la sélection de caractéristiques

A chaque variable d'optimisation, nous associons un bit de donnée. Nous définissons un chromosome comme une séquence de bits qui codent l'ensemble des variables de classification. Ce codage est effectué sur *66* bits pour les caractéristiques et *19* bits pour les descripteurs, regroupés en deux parties indépendantes et séparées. Ils indiquent l'activation ou la désactivation de chacune des *59* caractéristiques ou *12* descripteurs d'un graphème (premier *59* bits ou *12* bits) ainsi que la variation du seuil (*7* derniers bits) avec $S \in [3, 90]$. Un bit dont la valeur est "*1*" indique que la caractéristique correspondante est sélectionnée pour le clustering un «*0*» indique que cette variable n'est pas sélectionnée (figure 4.18 (a)).

3.3.1.2 Représentation du chromosome pour la pondération de caractéristiques

Pour chaque variable d'optimisation, nous associons un poids $w \in [0,1\text{-}0,9]$ où :

$$wi = Min + (Max - Min)\frac{F_{N=1\ldots\beta} - 1}{N - A} \times (0{,}85)^c \qquad \textbf{(4.6)}$$

- Avec *Min = 0,1* et *Max = 0,9*. A la caractéristique la moins discriminante sera affectée la valeur de poids *0,1*, cela nous garantit de n'avoir aucune valeur de caractéristiques nulle et de cette façon une caractéristique aura toujours un poids quelque soit son pouvoir discriminant. Les caractéristiques les plus discriminantes se verront affecter des valeurs élevées, la plus grande étant *0,9 (*au lieu de *1).* Cette approche de la pondération garantit ainsi la diversité des caractéristiques tout en valorisant les pouvoirs de discrimination élevés.
- Le terme $0,85^C$ permet d'exprimer la diversité des poids entre individus. Le scalaire *c* est l'indice de chaque individu dans une génération donnée. Ces poids sont ensuite normalisés tel que $\sum w_i=1$. Le codage du seuil reste le même pour la pondération.
- $F_{N=1..\beta}$: indice de la caractéristique
- $\beta = 59$ *ou 12* selon que l'on considère le nombre de caractéristiques ou le nombre de descripteurs (figure 4.18 (b)).

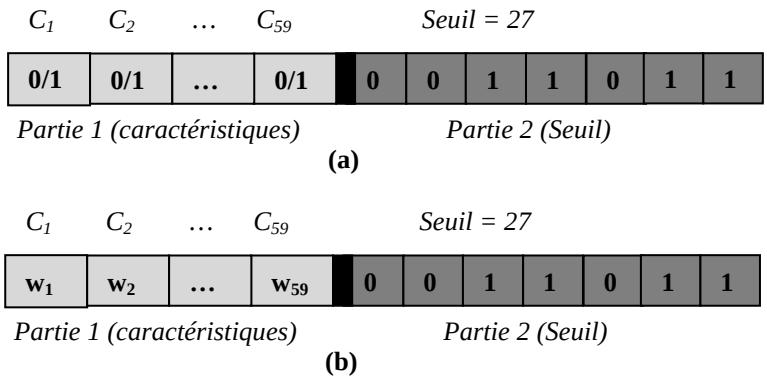


Figure 4.18. (a) Codage des paramètres de sélection en deux parties séparées d'un chromosome, (b) Codage des paramètres de pondération en deux parties séparées d'un chromosome

Le meilleur seuil utilisé par l'étape de coloration de graphes conduit à la séparation optimale des classes (nombre optimal de classes) et à une meilleure construction de dictionnaire de formes. Rappelons qu'une valeur de seuil faible conduit à une sous-segmentation des classes et une valeur élevée conduit à une sur-segmentation. Dans ces deux cas les taux de confusion et de fausse reconnaissance deviennent importants.

3.3.2 Initialisation et génération de la population initiale

Le choix de la population initiale (génération *0*) est important car il peut rendre plus ou moins rapide la convergence vers l'optimum global de la classification de graphèmes. Le démarrage est initié avec l'a priori fort qu'on ne connaît rien du problème à résoudre (pas d'a priori sur les résultats de classification). Nous avons donc réparti la population initiale sur tout

le domaine de recherche respectant la diversité entre les individus (caractéristiques et seuils). Cela permet de converger en des temps raisonnables et donc de ne privilégier aucune configuration de chromosomes. Chaque génération est donc composée de *50* chromosomes, avec une probabilité de croisement de *0,8* et une probabilité de mutation de *0,001*. Ces valeurs ont été fixées d'une manière expérimentale. Si deux chromosomes identiques existent dans la population initiale, un des deux sera regénéré.

3.3.3 Évaluation de la qualité de classification du dictionnaire de formes

Ce critère d'évaluation ne requiert aucune connaissance sur les résultats de classification des graphèmes à évaluer. Son principe consiste à estimer la qualité d'un résultat de classification à partir de statistiques calculées sur chaque classe formée. Cette mesure de qualité est établie en accord avec l'intuition humaine sur les conditions que devrait remplir une classification pour être considérée comme bonne. La qualité de la classification est calculée à partir de la combinaison entre l'uniformité intra-classe et la disparité inter-classes.

$$F = M_{inter\text{-}classes} + M_{intra\text{-}classes} \quad \textbf{(4.7)}$$

Ces deux critères sont inspirés de l'étude de *(Levine et Nazif, [1985])* sur la segmentation automatique des images naturelles en régions. Ils vont servir de mesures pour ajuster automatiquement le seuil d'adjacence, le sous-ensemble de caractéristiques optimal et évaluer la qualité de la classification. Nous les présentons formellement dans les sections suivantes :

3.3.3.1 Uniformité intra-classe

Le principe est ici de calculer l'uniformité d'une description d'un individu sur une classe en se basant sur la variance de cette description. La mesure d'uniformité $M_{intraclasses}$ issue d'une coloration de graphe en *k* classes est la suivante :

Nous définissons le barycentre m_i de la classe C_i par :

$$m_i = \frac{1}{card(C_i)} \sum_{vl \in Ci} v_l \quad \textbf{(4.8)}$$

Puis nous définissons la variance de la classe c_i par :

$$\mathrm{var}_i = \frac{1}{card(C_i)} \sum_{vj \in Ci} \left\| v_j - m_i \right\|^2 \quad \textbf{(4.9)}$$

La fonction $\| \ \|$ représente ici la norme classique euclidienne. La partie intra-classe de la fonction de fitness est définie par :

$$M_{intraclass} = 1 - \frac{1}{k} \sum_{i=1}^{k} \frac{\mathrm{var}_i}{\max_{vj \in Ci}(\left\| v_j - m_i \right\|^2)} \quad \textbf{(4.10)}$$

3.3.3.2 Disparité inter-classes

D'un point de vue complémentaire à l'uniformité intra-classe, un autre critère utilisé est la disparité inter-classes. En effet, deux classes voisines sont supposées avoir un contenu différent. Il devrait donc y avoir une disparité entre ces deux classes. Pour mesurer la disparité locale entre deux classes C_i et C_j, nous calculons la distance moyenne entre les caractéristiques des deux barycentres.

$$disp(C_i, C_j) = \frac{1}{59} \sum_{n=1}^{59} \left| m_i(n) - m_j(n) \right| \quad \textbf{(4.11)}$$

Où $m_i(n)$ est la valeur de la caractéristique n dans le barycentre de la classe C_i. Nous tenons à souligner que les caractéristiques sont déjà normalisées entre *0* et *1* (en utilisant la soustraction du *min* et la division par *(max-min)*.

Nous pouvons ainsi définir la disparité de classe C_i avec k classes par la formule suivante :

$$Disp(C_i) = \sum_{j=1 \ and \ i \neq j}^{k} disp(C_i, C_j) \quad \textbf{(4.12)}$$

La disparité globale entre les k classes est alors définie par la formule suivante :

$$M_{interclass} = \frac{\sum_{i=1}^{k} w_i \cdot Disp(C_i)}{\sum_{i=1}^{k} w_i} \quad \textbf{(4.13)}$$

Où w_i est un poids associé à chaque classe, et qui est lié à l'aire de la classe (le nombre de graphèmes dans une classe).

3.3.4 Sélection des individus, croisement et mutation

A chaque génération, les individus se reproduisent, survivent ou disparaissent de la population sous l'action d'un opérateur de sélection. Après avoir été classés en ordre décroissant en fonction de la fitness de chacun des individus de la population P_o, nous utilisons la technique de sélection par tournoi. Cette méthode donne des résultats plus satisfaisants que la méthode basée sur l'élitisme qui consiste à copier un ou plusieurs des meilleurs chromosomes dans la nouvelle génération. Ensuite, on génère le reste de la population selon l'algorithme de reproduction usuel. Cette méthode améliore considérablement les algorithmes génétiques, car elle permet de ne pas perdre les meilleures solutions. Dans la méthode de la roulette, les parents sont donc sélectionnés en fonction de leur performance. Meilleur est le résultat codé par un chromosome, plus grandes sont ses chances d'être sélectionné. Par le choix du tournoi, nous sélectionnons deux individus de la population qui sont nécessaires pour produire les enfants de la nouvelle génération par croisement à deux points sur chacune des parties, voir figure 10. Nous répétons cette procédure *m* fois pour obtenir les *n* individus de la nouvelle population $P_{o'}$, qui serviront de nouveaux parents. L'opérateur de croisement est conçu pour enrichir la diversité de la population en manipulant la structure des chromosomes. Le croisement est appliqué sur les deux parents (P_1 et P_2) et génère deux enfants (E_1 et E_2) (figure 4.19).

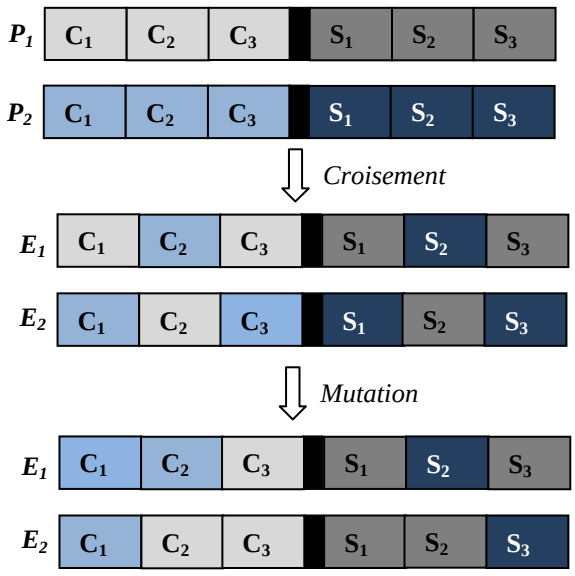


Figure 4.19. Principes de croisement et de mutation entre deux chromosomes

Par la suite, nous appliquons une mutation sur chacune des deux parties du chromosome (dans la pratique la probabilité de mutation utilisée est de *0,001*). L'opérateur de mutation a pour but de garantir l'exploration de l'espace d'états. Le processus s'arrête quand la fonction de

fitness devient stable ou lorsqu'un nombre maximal de générations (fixé à *1000*) est atteint. Si la fonction de fitness reste stable durant *50* générations, le processus de calcul des générations suivantes est arrêté. On conserve alors le sous ensemble de caractéristiques ayant conduit à la meilleure fonction de fitness.

4 Principe théorique de la coloration de graphes

La coloration de graphe constitue une branche très importante de la théorie de graphes. Ses applications sont nombreuses dans différents domaines scientifiques (optimisation des réseaux de transports ou de communication, des formules chimiques, …). Les définitions de la coloration sont simples et de véritables problèmes de recherche peuvent être posés sous une forme bien structurée dont la formulation peut recouvrir de grandes difficultés pratiques. Ce modèle a été introduit la première fois dans le domaine de la classification d'éléments de contenus connexes présents dans des images de documents d'entreprise par *(Gaceb et al. [2009]).* Les auteurs l'ont adapté à toutes les étapes d'analyse de ces documents (extraction de la structure physique et sa localisation à la reconnaissance) pour consolider la coopération et assouplir les échanges d'information entre les différents modules de segmentation et de reconnaissance. Grâce à sa simplicité et à son potentiel en matière de classification, nous avons pu imaginer une méthode originale de construction de dictionnaires de formes représentatifs de la distribution des graphèmes de l'écriture et de leurs fréquences d'apparition.

4.1 Modélisation du problème de classification de graphèmes en termes de coloration

Le regroupement d'un ensemble $X=\{x_1,..,\ x_n\}$ de n graphèmes en plusieurs groupes homogènes se base sur le principe que chaque groupe doit réunir le plus de graphèmes similaires. Les regroupements portent sur un critère de similarité S. Ce critère spécifie que certaines paires de graphèmes $\{x_i\ ,\ x_j\}$ ne peuvent être fusionnées au sein d'un même groupe. Pour résoudre ce problème de partitionnement (ou de classification), nous pouvons partir du point de vue inverse et formuler la question suivante « quel est le plus petit nombre de groupes homogènes que nous pouvons former en respectant la contrainte *S* (critère de similarité) ». L'intérêt de formuler le problème de cette manière, est qu'il devient alors possible de le modéliser en termes de coloration de graphe. Le positionnement du problème est alors le suivant : nous représentons chaque graphème x_i par un sommet $v_i \in V$ d'un graphe simple *G* et nous ajoutons une arête $E(v_i\ ,v_j)$ entre chaque paire de graphèmes dissemblables (qui ne respectent pas la contrainte *S*) (figure 4.20).

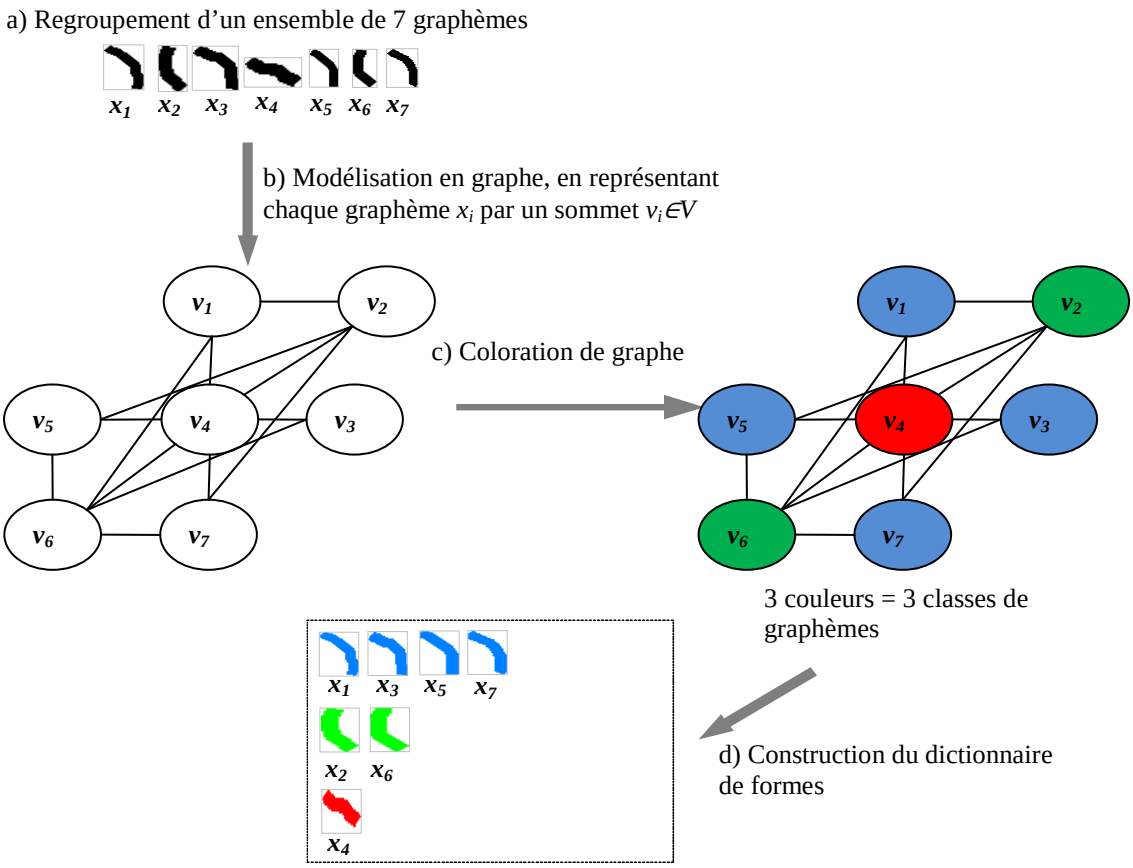


Figure 4.20. Etapes de construction du dictionnaire de formes par coloration de graphe

La coloration des sommets du graphe *G(V,E)* consiste alors à affecter à tous ses sommets une couleur de telle sorte que deux sommets adjacents (dissemblables) ne puissent pas porter la même couleur. Ces couleurs vont correspondre aux différents groupes homogènes qui constituent les différentes classes de graphèmes. Dans ce problème de regroupement, la question de la détermination du plus petit nombre de groupes homogènes, revient à rechercher le plus petit *k* pour lequel le graphe *G* correspondant admet une *k-coloration* : c'est donc précisément le nombre chromatique χ(G) du graphe *G* qu'il faut déterminer. De plus, cette modélisation présente l'avantage de gérer facilement plusieurs sortes d'ambiguïtés inhérentes à la forme des graphèmes par rapport aux mécanismes de regroupements classiques.

4.2 Construction du dictionnaire de formes

A partir de l'ensemble des graphèmes extraits selon la méthode de décomposition présentée dans le chapitre 3, nous procédons à la construction du dictionnaire de graphèmes des écritures traduisant les dissimilarités entre les formes.

4.2.1 Mesure de similarité

La dissimilarité entre deux graphèmes représentés par les sommets v_i et v_j dans le graphe est donnée par la distance généralisée de Minkowski d'ordre α ($\alpha = 2$: distance euclidienne).

$$D_s = \left(\sum_{k=1}^{n} g_k (v_i^k, v_j^k)^{\alpha} \right)^{\frac{1}{\alpha}} \quad \textbf{(4.14)}$$

$n=59$ est la longueur des vecteurs des caractéristiques. g_k est la fonction de dissemblance représentant la distance euclidienne qui compare les vecteurs de caractéristiques deux à deux.

4.2.2 Construction du graphe

La construction d'un graphe G à colorer à partir d'un ensemble $X=\{x_1,\ldots,x_n\}$ de n graphèmes (où chaque sommet v_i correspond au vecteur descripteur du graphème x_i) est principalement basée sur le calcul de la matrice de distances MD_s . Cette matrice traduit les dissimilarités $Ds(x_i, x_j)$ entre les paires de graphèmes (x_i, x_j) données par la relation suivante : $MDs[v_i,v_j]=Ds(x_i, x_j)$ avec $i \in [1,n]$ et $j \in [1,n] \mid (i \neq j)$. Une fois MD_S calculée, nous associons à X un graphe seuil supérieur $G_{\geq S} = (V=X, E_{\geq S})$ en utilisant la relation suivante :

$$E(v_i, v_j) \in E_{\geq S}[vi, vj] \text{ si } Ds(x_i, x_j) = MDs(v_i, v_j) \geq S \quad \textbf{(4.15)}$$

Remarquons que le terme adjacence (ou voisinage) est différent du terme similarité. En effet, deux sommets sont adjacents s'ils ont une dissimilarité supérieure au seuil S. Le seuil S est également nommé seuil d'adjacence. Ce seuil peut être ajusté manuellement à l'aide des paléographes ou automatiquement en maximisant la qualité de classification ψ donnée par (Gaceb et al. [2008]), où:

$$S^{optimal} = \{\arg\max(\psi(S_i))\} \quad \textbf{(4.16)}$$

Dans les sections suivantes les comparaisons tous les graphes qui présentent les résultats des algorithmes génétiques montrent les moyennes de la fonction de fitness.

5 Comparaison de la sélection et pondération

Les AG présentent l'intérêt de pouvoir être exploités soit pour une sélection de caractéristiques, soit pour leur pondération directe (en attribuant des valeurs de poids non nulle à chaque caractéristique). Dans cette section nous proposons donc de comparer l'approche de sélection de caractéristiques avec celle de la pondération, le but de cette comparaison est de montrer la supériorité de la méthode de pondération sur celle de la sélection des

caractéristiques. L'évaluation est appliquée sur une base de *10000* graphèmes issus de la décomposition de *20* pages de manuscrits de la base d'Oxford de différents styles. Nous avons comparé les résultats des fonctions fitness entre d'une part les résultats de classement selon le mécanisme de pondération de caractéristiques et d'autre part selon le mécanisme plus simple de sélection de caractéristiques sur les *20* pages de manuscrits. La figure 4.21 montre que la classification des graphèmes selon le mécanisme de pondération de caractéristiques (portant sur *59* valeurs) et de pondération des descripteurs (portant sur les groupements de *11* valeurs) produisait des résultats supérieurs aux autres méthodes de sélection (sélection des descripteurs et sélection des caractéristiques) avec une fitness maximale de *0,990* pour la pondération de caractéristiques et une fitness de *0,9890* pour la pondération de descripteurs. Cela est dû au fait que la sélection de caractéristiques (resp. de descripteurs) peut éliminer des caractéristiques (resp. des descripteurs) discriminants ce qui aura un effet négatif sur le résultat de classification, tandis que la pondération attribuera un poids peu important aux caractéristiques (resp. aux descripteurs) non discriminants et un poids plus élevé aux caractéristiques (resp. aux descripteurs) plus importants. De cette façon, toutes les caractéristiques (les *59* valeurs) participent avec une pondération spécifique à la classification.

Rappelons ici que les mesures de Fitness ont été obtenues à partir d'un calcul de moyenne sur *10* passages successifs par le processus complet de sélection/pondération comme cela doit se produire pour une estimation réaliste lorsque les AG sont impliqués.

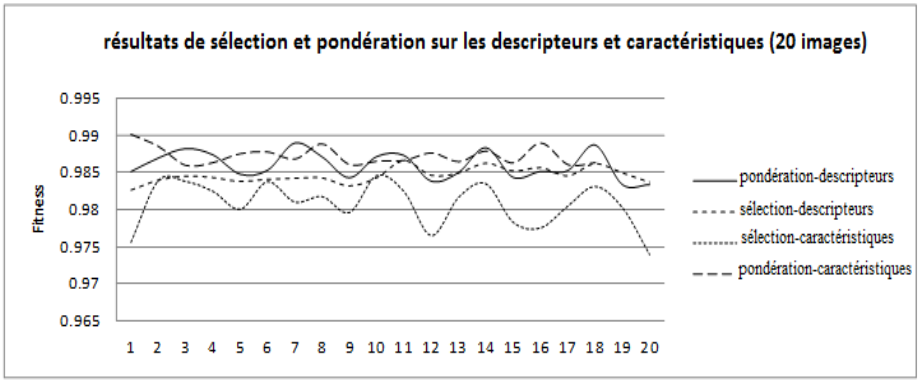


Figure 4.21. Courbes représentant les fitness maximales sur les 20 images en utilisant les techniques expliquées dans les sections précédentes

Pour le même manuscrit (figure 4.22) nous pouvons observer l'évolution de la fonction de fitness et du seuil de coloration pour la pondération de caractéristiques. Dans cet exemple, la stabilité de la fonction de fitness pour une valeur de *0,9887* est atteinte au bout de *47* générations. Comme elle devient stable pour les *50* générations suivantes le processus de création de nouvelles générations s'arrête et l'individu avec la meilleure fonction de fitness est choisi. On remarque également que le seuil de coloration n'est pas stable durant les *47*

premières générations, cela signifie que l'AG poursuit son effort d'optimisation de ce seuil tout au long du processus de création des nouvelles générations en même temps que son optimisation dans la sélection du meilleur sous-ensemble de caractéristiques. La stabilité de la fonction de fitness est obtenue pour la valeur du seuil de *59* et un nombre de classes de graphèmes atteignant la valeur *10*.

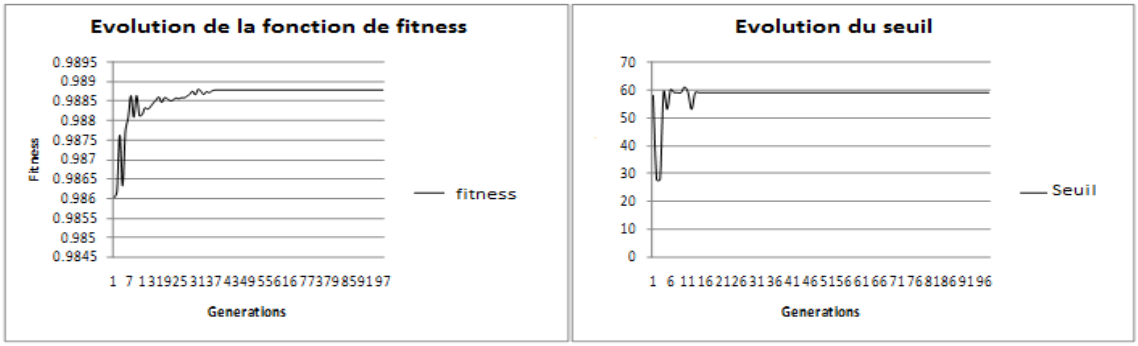


Figure 4.22. Evolution de la fonction de fitness et du seuil de coloration

Nous avons également étudié la distribution des poids attribués à chacune des 59 caractéristiques (figure 4.23) afin de voir si certaines d'entre elles possèdent une capacité discriminatoire supérieure aux autres. Les poids les plus élevés ont été attribués aux descripteurs suivants : *{D_9 = 9 Densités, D_{10}= Huit orientations, D_{11}=Les 9 directions, D_{12} = les Moments de Zernike}*, soient les caractéristiques correspondant aux bits numérotés *[9..59]*. La méthode de pondération attribue des poids plus élevés aux descripteurs de taille *1* (descripteurs sont décomposés sur une grille *3×3*). Notons également que les *4* derniers descripteurs ont été fortement pondérés également. Nous pouvons constater suite aux différentes expériences que nous avons menées sur les manuscrits de la base d'Oxford que les descripteurs *{D_1 = Hauteur, D_2 = Largeur, D_3 = excentricité}* et leurs caractéristiques ont été pondérés très faiblement. Pour ce type d'images d'écritures, cela montre qu'ils n'ont pas de pouvoir discriminant très fort (figure 23).

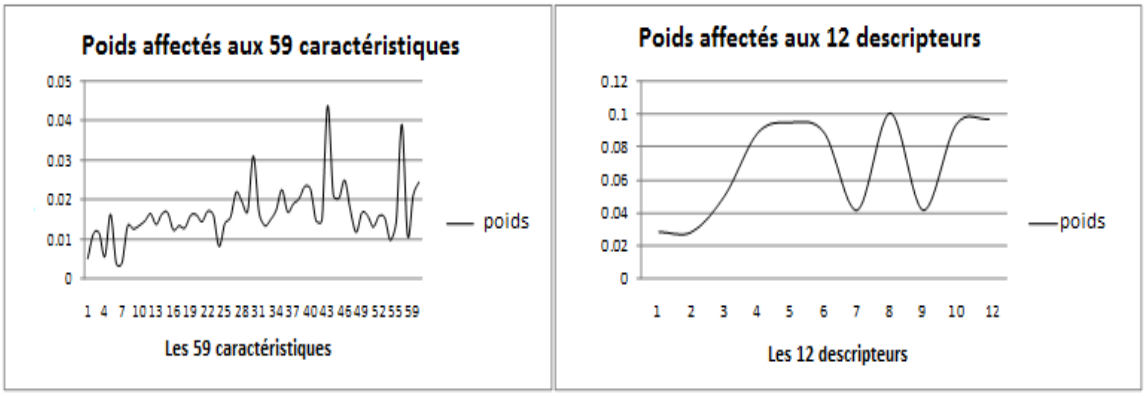


Figure 4.23. (a)Relation entre poids et caractéristiques, (b) poids et descripteurs

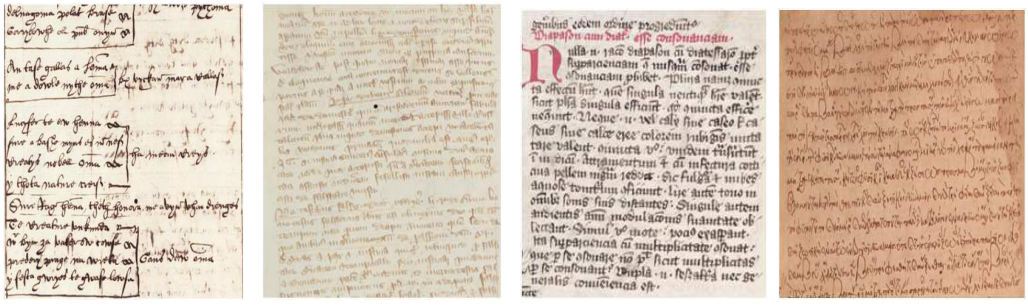
L'ensemble de ces résultats dont nous avons illustré les moyennes sur *10* expériences à chaque fois, nous permettent de mettre en avant le fait que la pondération de caractéristiques fournissait de meilleurs résultats de classification que la pondération de descripteurs et les méthodes de sélection. Nous avons donc pris le parti de nous intéresser au mécanisme de pondération, en construisant une pondération dite *générique*, que nous appliquerons pour la construction des dictionnaires de formes et qui permettra de n'écarter aucune caractéristique, privilégiant pour celles qui sont faiblement discriminantes des pondérations faibles.

6 Résultats et application

Nous avons l'évaluation des résultats de notre proposition de caractérisation des graphèmes en trois étapes :

1. Dans la première nous validons notre approche de sélection de caractéristiques (par un apprentissage non-supervisé) et nous montrons comment déduire un jeu de poids génériques sur la base d'apprentissage.
2. Dans la seconde étape nous démontrons la pertinence de ces poids sur une autre base (de test) qui se compose de manuscrits inconnus.
3. Dans la troisième étape nous associons à chaque style un dictionnaire de formes représentatif, et calculons les poids génériques spécifiques à chaque style.

Nous utilisons pour les trois étapes *140* images de manuscrits médiévaux de la base d'Oxford que nous avons décomposées en *600000* graphèmes. Nous avons ensuite divisé cette base en deux parties: la base *B1* comprenant *100* images pour l'étape d'apprentissage et la déduction des poids génériques et la base *B2* contenant *40* images pour l'étape de test et la validation des poids génériques (figure 4.6). Chaque page manuscrite est représentée par un ensemble de graphèmes (figure 4.24).



Style 1 Style 2 Style 3 Style 4

Figure 4.24. Les quatre styles présents dans la base d'Oxford

6.1 De la sélection de caractéristiques à la déduction des poids génériques

Dans cette première étape, nous sélectionnons avec l'AG (voir la section 3) *pour chaque manuscrit* de la base *B1* un sous-ensemble de caractéristiques les plus pertinentes parmi l'ensemble de *59* caractéristiques de départ $\{a_1 ... a_{59}\}$. Chaque manuscrit conduit à une sélection de caractéristiques propre (nous aurons donc *100* sélections portant sur les 100 manuscrits de la base B1). Puis nous associons un poids à chaque caractéristique a_i, calculé à partir des *100* sélections de la façon suivante: $w(a_i) = N(a_i)/T$; où $N(a_i)$ est le nombre de fois qu'une caractéristique a_i était sélectionnée sur les *n* manuscrits de la base *B1*.

$T = \sum_{i=1}^{59} N(a_i)$ est un paramètre de normalisation des poids. La somme de *59* poids est égale à *1*.

D'autre part, le seuil optimal moyen de coloration est ensuite calculé par *AG* pour chaque manuscrit de la base *B1*. Nous calculons ce seuil générique de coloration comme la moyenne de tous les seuils optimaux obtenus sur les *n* manuscrits avec:

$$S_{générique} = \sum_{i=1}^{n} S_i^{optimal} / n \qquad \textbf{(4.17)}$$

Les poids et le seuil génériques obtenus dans cette première étape sont utilisés pour calculer la nouvelle fonction de fitness sur chacun des manuscrits de *B1*. Nous comparons ensuite ces dernières mesures de fitness avec les fonctions de fitness maximales obtenues à la suite de chacune des sélections de caractéristiques. Cette comparaison va nous permettre de vérifier, pour chaque manuscrit, si les poids et le seuil génériques donnent de meilleures valeurs de finesses (ou proches) que celles données par la seule sélection de caractéristiques (Figure 4.25).

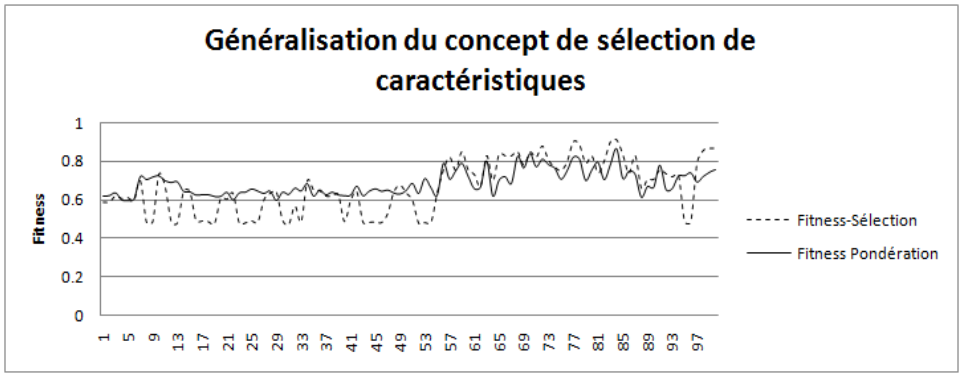


Figure 4.25. Courbes représentant les valeurs de la fonction de fitness qui sont calculées à partir des poids génériques et à partir de la sélection de caractéristiques sur les *n* (*n=100*) manuscrits de la base d'apprentissage

A partir de la courbe précédente, nous constatons qu'avec un seul jeu de poids et seuil génériques nous obtenons sur tous les manuscrits une fitness très proche ou meilleure que celle obtenue par les différentes sélections de caractéristiques. Ce résultat montre la possibilité d'appliquer ces poids et seuil génériques sur de nouveaux manuscrits (inconnus) en assurant une meilleure qualité de classification et par conséquent un dictionnaire plus représentatif des classes de graphèmes. C'est ce constat de faisabilité de la généralisation de ces poids et seuil que nous allons démontrer à travers l'expérience suivante.

6.2 Validation des poids génériques pour toutes les classes

Dans cette seconde étape, nous appliquons les poids et le seuil génériques issus de la première étape, cette fois-ci, sur les manuscrits de la base *B2*. La courbe suivante montre pour chaque manuscrit considéré sur l'axe des *x*, la valeur des différentes fonctions de fitness calculées dans trois contextes (*10* mesures successives ont été prises pour chaque manuscrit et moyennées ici sur la courbe). Nous considérons que les résultats vérifient l'hypothèse de la validation des poids génériques montrée dans la figure 4.6, en effet la courbe montre que la fitness des poids généralisés F^* est proche de *F'*, la fitness résultant de la sélection des caractéristiques ($F^* \approx F'$) et plus grande que *F* la fitness sans sélection ni pondération ($F^* > F$) (figure 4.26). Ainsi, nous pouvons utiliser les poids génériques et le seuil optimal sur un manuscrit inconnu que nous désirons introduire dans la base sans avoir recours à une sélection de caractéristiques par AG sur ce dernier.

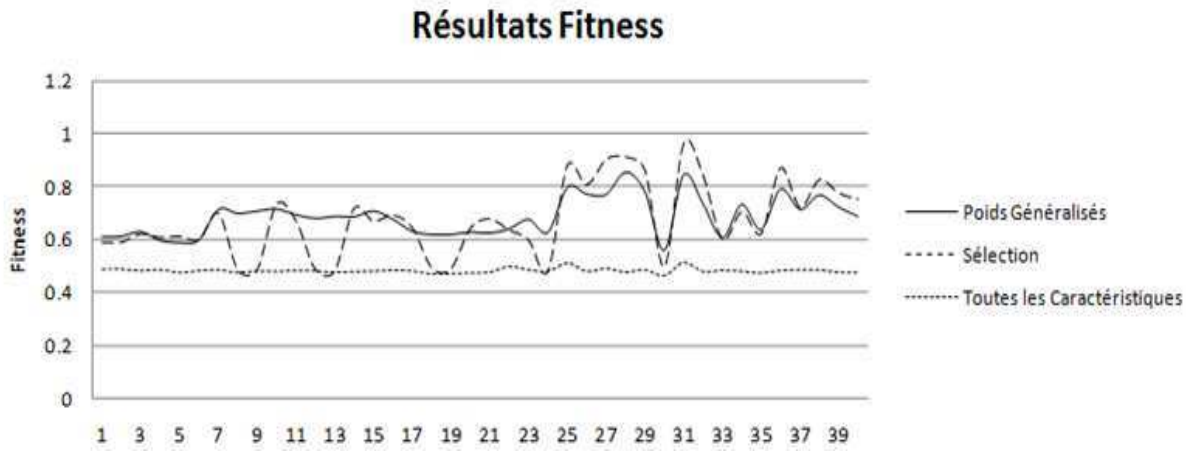


Figure 4.26. Résultats des valeurs de fitness à l'issue **des 3 tests**

Nous pouvons également constater à partir de la courbe suivante que les poids les plus élevés correspondent aux caractéristiques appartenant aux descripteurs suivants : D_9 = *9 Densités,* D_{10}= *Huit orientations,* D_{11}=*Moments de Zernike*. Cela signifie que ces dernières ont un pouvoir de discrimination supérieur aux autres (figure 4.27).

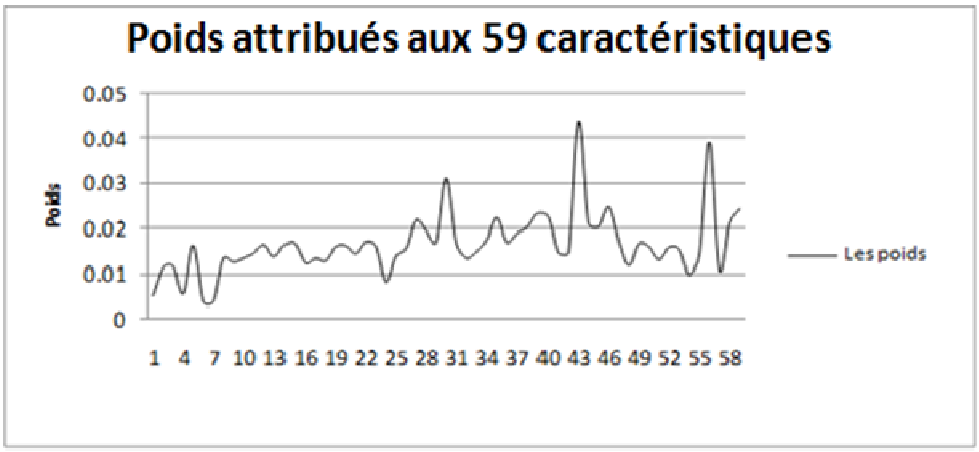


Figure 4.27. Relation entre poids et caractéristiques

Dans le chapitre suivant, nous allons voir comment appliquer les poids et le seuil génériques d'une manière efficace pour la reconnaissance de style par la technique *CBIR*. Pour faire une comparaison des images par le contenu, chaque manuscrit est décrit par son propre dictionnaire de formes. Nous rappelons qu'un dictionnaire de formes est construit par une coloration des graphèmes en utilisant les poids et le seuil génériques de l'étape d'apprentissage.

La figure 28 montre deux exemples de dictionnaires de formes. La première colonne à gauche, présente un exemple de manuscrit de la base Oxford colonne (figure 4.28 (a)), avec *118* classes de graphèmes après l'application des poids génériques et seuil d'adjacence $S = 15$ (figure 4.28 (b)). Nous remarquons que les graphèmes ayant des formes similaires se regroupent dans la même classe et que deux formes identiques mais ayant subi une rotation apparaissent dans deux classes distinctes, ce qui était initialement souhaité (figure 4.28 (c)). La deuxième colonne à droite, montre un dictionnaire de formes après l'application des poids génériques et le seuil optimal $S = 15$. Le dictionnaire de formes se compose de *103* classes de graphèmes associés à un manuscrit de la base de données IRHT.

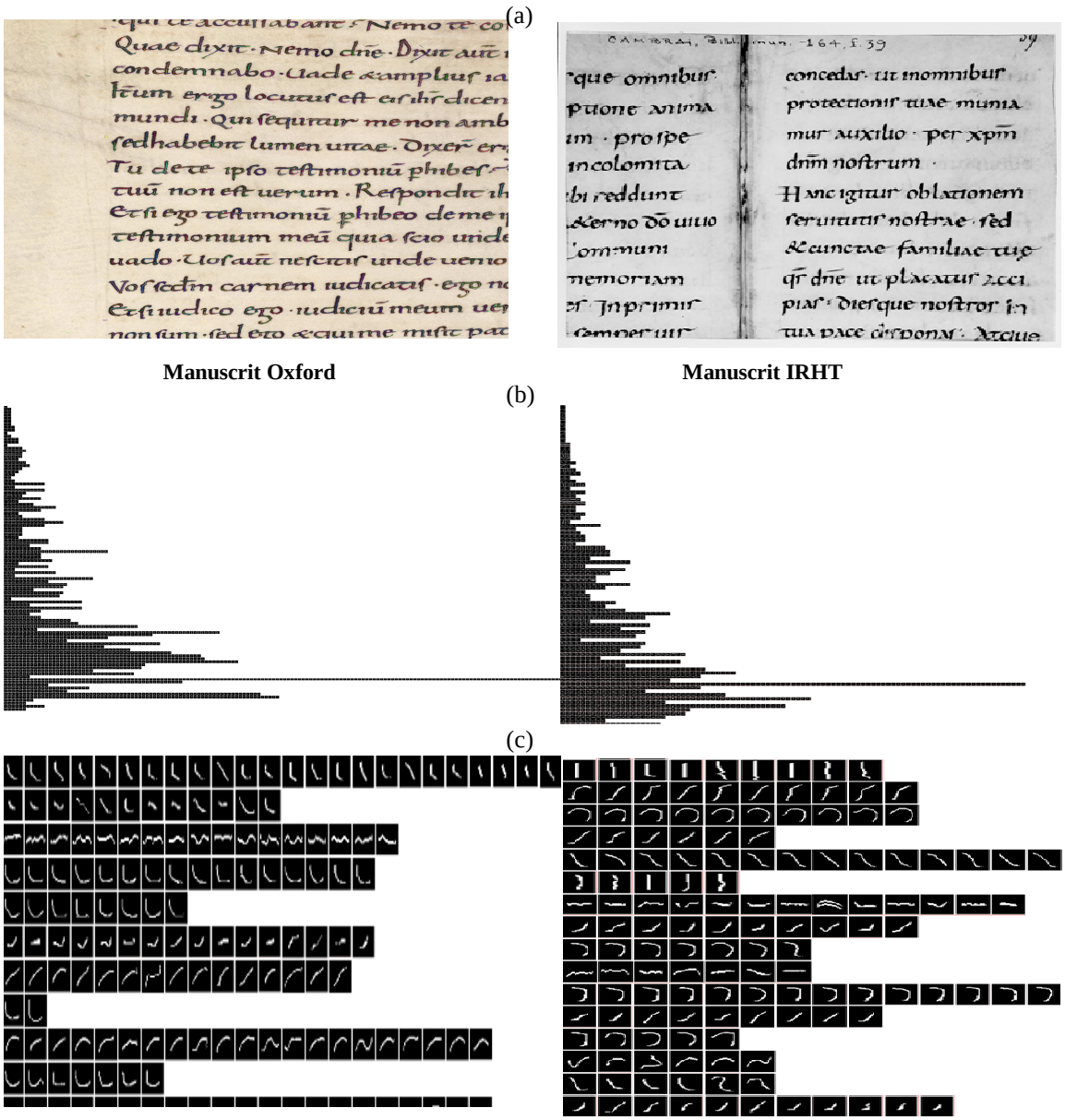


Figure 4.28. Résultat de construction des dictionnaires de formes utilisant la coloration de graphes et la pondération de caractéristiques sur des manuscrits de la base Oxford et IRHT

6.3 Poids génériques et dictionnaire de formes associés à un ensemble de documents

Dans cette section nous présentons notre méthode de sélection de dictionnaire de formes, de génération de poids génériques et seuil de coloration spécifique à chaque style. Le but de ce test est de savoir quel type d'approche choisir pour générer des poids génériques et un seuil optimal sur l'ensemble de la base : avoir recours à un dictionnaire de formes représentatif de chaque style ou disposer d'une pondération générique unique pour l'ensemble de la base.

6.3.1 Dictionnaire de formes représentatif

Pour chaque style $S_{i=1...4}$ appartenant à la base *B1*, et composé de *ni* dictionnaires de formes *(*$\sum ni$ *=100)* $cb(p_1)...cb(p_{ni})$ nous cherchons le dictionnaire de formes représentatif qui représente le centre du style.

Pour un ensemble de de dictionnaires de formes C(b) appartenant à un style S_i ***:***
Pour chaque dictionnaire de formes $Cb(p_{k=1...ni})$ *:*
Somme_des_distances = 0
  Pour chaque dictionnaire de formes $Cb(p_{t=1...ni})$ *:*
    Si ($Cb(p_t) == Cb(p_k)$) *{continue}*
    Sinon
      Somme_des_distances += $d_p[cb(p_t), cb(p_k)]$ (Section 6.3).
  Si (Somme_des_distances < δ) alors
    δ = Somme_des_distances ;
    Centre_style = $cb(p_t)$;
  Fin Si

La distance de chaque dictionnaire de formes $cb(p_t) \in S_i$ est calculée par rapport à tous les autres dictionnaires de formes $cb(p_k)$ $(k \neq t)$ du même style S_i . Le centre d'un style S_i est celui qui produit la plus petite distance.

6.3.2 Validation du dictionnaire de formes représentatif, des poids génériques et du seuil spécifique à chaque style

La (figure 4.29 (a)) présente les valeurs des poids génériques de chacune des *59* caractéristiques calculées sur toute la base. Et la (figure 4.29 (a)) présente les valeurs des poids génériques obtenus sur chacun des quatre styles.

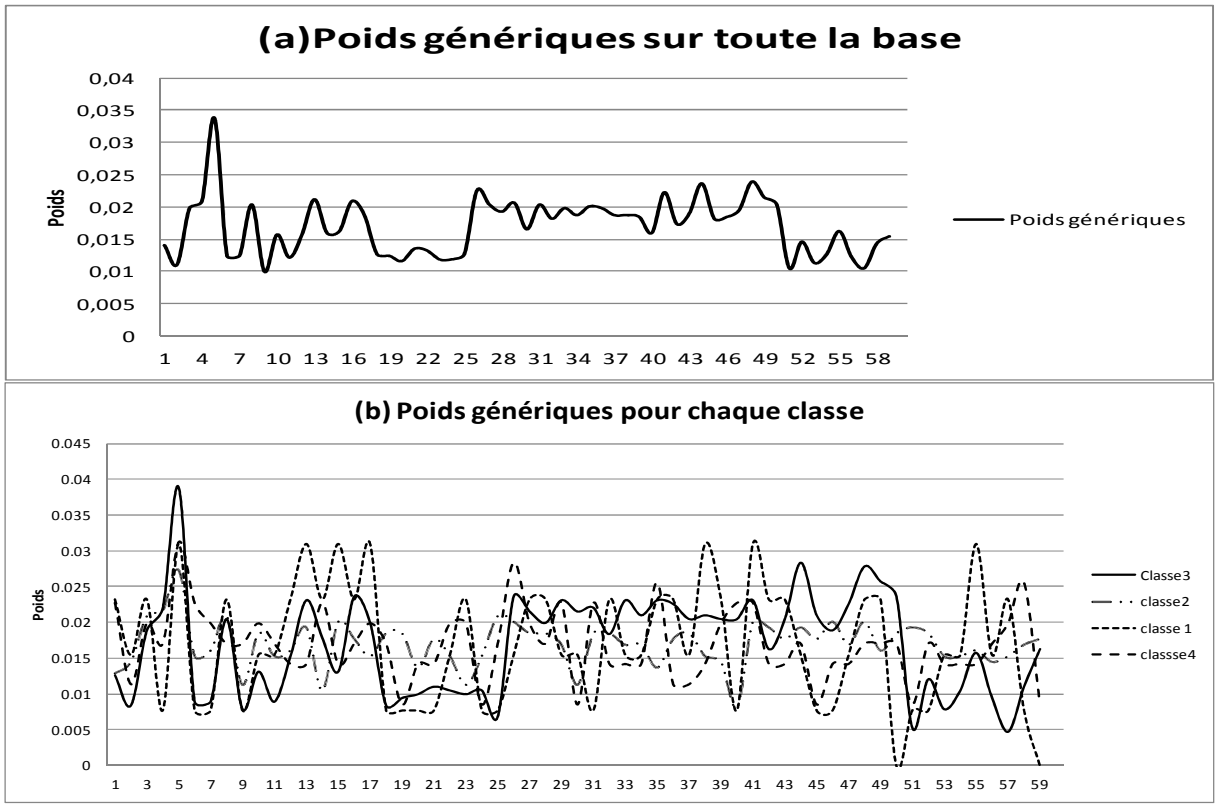


Figure 4.29. (a) distribution des poids pour toutes les classes de la base B1, (b) distribution des poids pour chaque classe de la base B1

Le tableau 4.4 représente la distance euclidienne calculée entre les poids génériques de chacune des quatre classes de styles et normalisée entre *1* et *10*. On remarque que les *styles 1*, 3 et *4* dans *B1* possèdent des propriétés de formes semblables et très différentes de celles de style *2*. Cela signifie que les distances entre les poids dans des classes de styles proches sont faibles.

Tableau 4.4. Distances entre les poids des classes

	Classe *1*	**Classe *2***	**Classe *3***	**Classe *4***
Classe *1*	*0,0*			
Classe *2*	*6,0*	*0,0*		
Classe *3*	*3,1*	*4,5*	*0,0*	
Classe *4*	*4,8*	*4,0*	*1,2*	*0,0*

L'utilisation de dictionnaire de formes représentatif pour chaque classe de styles sera justifiée dans le chapitre 5. Nous montrerons notamment qu'il n'est pas optimal d'exploiter une telle approche (un discitonnaire par style) sur de petites bases de manuscrits. En revanche, cela n'est plus vrai pour les grandes bases de plusieurs dizaines de styles (voir des centaines de styles, que nous n'avons pas ici dans notre étude). Ces dictionnaires de formes représentatifs ne seront pas exploités dans notre étude car comme le tableau suivant le montre les différences ne sont pas suffisamment significatives.

7 Conclusion

Dans ce chapitre, nous avons montré, à partir de la décomposition en graphèmes, qu'il était possible de produire une description des formes (par un ensemble de *59* caractéristiques rassemblés en *12* descripteurs) directement interprétables par un expert humain. L'analyse locale ainsi produite correspond à une approche de la description des formes qui permet une interprétation par l'expert et conserve une grande pertinence pour l'approche de classification par coloration de graphe qui a été mise en œuvre. La formation des dictionnaires de formes propres à chaque manuscrit est l'expression de la signature de son style. Cette signature est unique et permet de réaliser des rapprochements entre manuscrits en les comparants selon des métriques adaptées.

Nous avons montré l'importance de l'algorithme génétique (AG) combiné à la méthode de coloration de graphes pour la construction des dictionnaires de formes. Cette combinaison nous permet non seulement de sélectionner un sous-ensemble de caractéristiques qui donne de bons résultats de classification, mais également de comprendre l'impact et le pouvoir discriminant de chaque caractéristique. Nous avons également proposé une approche générique permettant de monter qu'à partir des poids génériques spécifiques à chaque style nous pouvions analyser les propriétés communes des manuscrits de chaque style d'écriture et vérifier leurs similarités internes. Ces différentes propositions seront évaluées dans le chapitre applicatif suivant. Nous montrerons notamment comment l'aspect hybride de notre approche de caractérisation (basée sur une analyse locale et globale en même temps) pouvait offrir de meilleurs taux de reconnaissance de styles par rapport aux méthodes existantes, à travers une application de type CBIR. Dans les perspectives, nous considérerons également l'idée de déduire des poids génétiques à partir de la méthode de pondération directe issue des AG. Au lieu de sélectionner les caractéristiques et puis les moyenner, nous pouvons générer des poids en utilisant la méthode présentée dans la section 3.2.1.1 puis moyenner les poids. L'ensemble de ces expériences sont décrites dans le chapitre 5.

Chapitre 5 : Exploitation des dictionnaires de formes pour la comparaison des écritures

Résumé: *Dans ce chapitre nous exploitons les dictionnaires de formes des manuscrits pour la classification de styles et l'identification de scripteur de différents types de manuscrits qui s'étendent de l'époque carolingienne jusqu'à nos jours. La similarité de deux dictionnaires de formes composés de clusters de graphèmes est faite en calculant une distance de similarité entre les clusters. Dans cette étude nous comparons la distance Hausdorff, la distance de Schaefer qui est une modification de la distance de Hausdorff et la distance Quadratique. Les résultats sur les bases IRHT, Oxford et ICDAR 2011, montrent la pertinence de notre méthode hybride qui combine les dictionnaires de formes représentant une signature globale d'un manuscrit et les caractéristiques locales extraites à partir des graphèmse. Elle classifie les styles et identifie les scripteurs. Les résultats montrent aussi que notre méthode peut être généralisée sur tout type de style d'écriture, elle n'est pas spécifique à un style ou une époque.*

Mots clés : *dictionnaire de formes, mesure de similarité, CBIR, reconnaissance de style, identification de scripteur*

1 Introduction

Après avoir décomposé les manuscrits en graphèmes, caractérisés ces derniers et construit des dictionnaires de formes, nous passons à la dernière partie qui consiste à exploiter ces dictionnaires de formes spécifiques à chaque manuscrit dans deux domaines : classification des styles et d'identifications des scripteurs.

La classification des manuscrits selon le style d'écriture et l'identification de scripteurs sont devenus des domaines de recherche très importants. Pour la classification des écritures, des représentations invariantes sont recherchées et sont capables d'éliminer les variations entres les différentes écritures. Par contre le problème d'identification du scripteur, exige une amélioration spécifique de ces variations, celles qui sont spécifiques à un scripteur. La classification de l'écriture par style et l'identification du scripteur représentent donc deux facettes opposées de l'analyse de l'écriture *(Bulacu et al. [2007]).*

Dans ce chapitre nous exploitons tout d'abord notre méthode afin d'aider les paléographes à classifier les manuscrits médiévaux par style d'écriture, ce qui présente l'objectif principal de nos travaux. Puis nous testons la pertinence de notre méthode à pouvoir prendre en compte des variations de l'écriture pour identifier les scripteurs. Dans les deux domaines nous allons

utiliser les dictionnaires de formes qui ont montré une grande efficacité en ce qui concerne les résultats de reconnaissance *(Schomaker et al. [2004])*. La représentation des dictionnaires de formes considérés comme signatures d'un objet (manuscrit) et les méthodes de calcul de distance de similarités entre ces signatures ont une influence sur les résultats de classification et identification comme le montre les résultats dans *(Bulacu et Schomaker, [2006])*. La notion de similarité devient ainsi centrale.

Après avoir redéfini de façon formelle la notion centrale de dictionnaires de formes, nous présenterons un état de l'art sur les mesures et distances qui sera suivi des expériences que nous avons faites sur les trois bases de notre étude (base de manuscrits médiévaux de l'IRHT, base de manuscrits médiévaux d'Oxford et base de manuscrits contemporains de la compétition ICDAR 2011. Nous terminerons le chapitre par une analyse et un comparatif des résultats.

2 Mesures de similarité

La notion de similarité, dont nous n'allons présenter que quelques aspects dans le texte qui suit, prend son origine dans la perception visuelle humaine. Nous commencerons par un rapide historique de cette notion de similarité d'un point de vue perceptuel. Les premières études ont établi quelques propriétés et quelques règles que l'on peut illustrer par quelques exemples concrets. Cette notion de similarité a ensuite été formalisée de façon plus précise pour conduire à une évaluation sous la forme d'une mesure le plus souvent. Les expressions «mesure de dissimilarité» et «mesure de similarité» souvent rencontrées dans la littérature ne seront pas distinguées dans ce rapport. Dans les deux cas, la mesure traduit soit ce qui est similaire, soit ce qui est dissimilaire entre deux objets (images ou vecteurs). Sous le terme de dissimilarité (sans doute moins employé que similarité), est donc sous-entendue une mesure d'autant plus grande que les objets comparés sont différents.

2.1 Un peu d'histoire

L'histoire de la notion de similarité d'un point de vue scientifique commence avec les études réalisées par Gustav Fechner (1860) pour relier un stimulus physique à une impression, une perception, représentable ou projetable sur une échelle de mesure. La similarité se traduit alors par une impression (ou mesure) identique. Ce modèle de Fechner présente cependant des lacunes de répétabilité d'une stimulation à une autre, liées à l'aspect très subjectif de cette évaluation. Entre deux stimulations, les modifications de contexte (ou l'effet de la première stimulation) changent le ressenti. Les travaux suivants de Louis Léon Thurstone (1927) proposent une autre approche, plus directement basée sur la comparaison de deux ou plusieurs stimuli simultanés. La question posée concerne des comparaisons de deux ou plusieurs stimuli.

Ces premiers travaux portent sur des expériences dont les propriétés psychologiques sous-jacentes sont modélisables en une dimension. Les travaux suivants porteront sur des stimuli plus complexes, qui contiennent plusieurs caractéristiques ou facteurs. Le sujet de l'expérimentation doit donner les pairs les plus semblables et les plus dissemblables parmi plusieurs stimuli complexes. Dans sa forme originelle, cette expérience est une extension naturelle de l'approche monodimensionelle de la similarité développée par Fechner puis Thurstone. On parlera alors de stimuli multidimensionels.

Avec ces stimuli multidimensionels, la notion de similarité revient à l'évaluation du nombre de caractéristiques communes (ou suffisamment proches) entre les stimuli. Cette opération de comparaison entre caractéristiques n'est pas encore abordée. Elle sera centrale dans la partie suivante. Pour l'instant, c'est le sujet humain qui est seul juge pour la comparaison de chaque caractéristique. La construction d'une similarité plus globale issue de la fusion de la similarité évaluée sur chaque caractéristique est l'objet des travaux de (Tversky, [1977]). Dans son modèle des contrastes, il donne quelques propositions pour cette comparaison. D'après lui, la similarité entre deux stimuli qui présentent de multiples caractéristiques est calculée par le sujet en prenant en compte les caractéristiques que les stimuli ont en commun et celles qui les différencient. Il faut donc :

- isoler les caractéristiques significatives de chaque stimulus.
- calculer le nombre de caractéristiques communes aux deux stimuli.
- calculer le nombre de caractéristiques associées à un stimulus mais pas à l'autre et vice versa.
- affecter un poids à l'ensemble des caractéristiques qui sont communes et celles qui sont différentes en fonction de leur importance.
- soustraire ces deux ensembles pondérés pour obtenir un indice de similarité.

L'une des propriétés importantes de ce modèle des contrastes de Tversky est rapportée par (Tversky et Gati [1978]) et concerne la propriété de similarité asymétrique, qui ne se retrouve que dans très peu d'autres modèles. La similarité asymétrique se traduit par le fait qu'un sujet humain à qui on demande de comparer deux stimuli *A* et *B* peut donner une estimation de ressemblance entre *A* et *B* différente de la ressemblance entre *B* et *A*. L'une des premières expériences mettant en évidence cette propriété portait sur les similarités entre pays. Ainsi, la Pologne était jugée en moyenne plus ressemblante à la Russie, que la Russie à la Pologne. De même, le Luxembourg était jugé en moyenne plus ressemblant à la Belgique, que la Belgique au Luxembourg. Cela peut s'analyser de la façon suivante : le stimuli *A* qui possède pratiquement

toutes ses caractéristiques similaires à *B* est jugé très similaire à *B*. En revanche *B* pourra se différencier perceptuellement de *A* si il possède des caractéristiques qui lui sont spécifiques.

Toutes ces expériences psychologiques ainsi menées, permettent de déterminer une notion du type « distance » entre les différents stimuli présentés. Cependant, cette notion de distance ne donne pas une information de mesure sur une échelle ou un axe, ce qui est un point essentiel pour une analyse de ces stimuli. Le passage de la notion de distance (relative) entre stimuli à la notion de mesure (plus absolue) associée aux caractéristiques des stimuli correspond à l'étape suivante de l'évolution d'une similarité perceptuelle vers une similarité plus objective, plus quantifiable. Au sens plus strictement mathématique, le concept de similarité correspond alors à la représentation des caractéristiques d'un stimulus sous la forme d'un vecteur (de caractéristiques) constitué de valeurs réelles.

Si on suppose que la notion de similarité étudiée peut se relier à la notion de proximité de vecteurs dans un espace qui hérite de propriétés de métrique, alors on peut utiliser les techniques de « multidimensional scaling » (MDS). Elles assimilent les stimuli complexes précédents à des points dans un espace muni d'un repère composé d'axes de références. Le problème des techniques de « multidimensional scaling » commence donc par une modélisation. Il faut répondre aux questions suivantes :

- Les stimuli peuvent-ils ou non être représentés par des points dans un espace (Euclidien ou pas …) ?
- Si oui, quel est le nombre minimum de dimensions de l'espace en question ?
- Comment réalise-t-on les projections de ces points sur les axes de références ainsi déterminés.

De façon très synthétique, les techniques de MDS présentent donc deux étapes : la première consiste à transformer les notions de taux ou niveaux de similarités en distances, la seconde détermine le positionnement des points dans un espace de dimension minimum. Deux caractéristiques essentielles doivent être associées à ces espaces :

- Ils servent de modèles psychologiques. A ce titre, ils doivent traduire le plus fidèlement possible la notion de similarité. En particulier, deux stimuli similaires doivent partager un même voisinage (notion de topologie en relation avec la notion de distance).
- Ce sont des espaces métriques. A ce titre, les distances calculées dans cet espace vérifient les propriétés classiques qui définissent les distances.

Ces propriétés qui définissent une distance d entre stimuli sont les suivantes :

- Pour tout stimulus *S*, $d(S,S) = 0$.
- Pour tous stimuli *SA* et *SB*, $d(SA,SA) \leq d(SA,SB)$.
- Pour tous stimuli *SA* et *SB*, $d(SA,SB) = d(SB,SA)$ (Symétrie).
- Pour tous stimuli *SA*, *SB*, et *SC*, $d(SA,SC) \leq d(SA,SB) + d(SB,SC)$ (Inégalité triangulaire).

Après avoir présenté l'évolution de la notion très générale de similarité entre stimuli et les études qui ont été menées pour passer d'une notion perceptuelle (difficilement accessible à la mesure) à une notion plus quantifiable, nous allons nous focaliser sur des stimuli plus spécifiques : les images. L'ensemble de ce que nous avons présenté sur des stimuli très généraux s'applique aux images, et à leur contenu. De même, les modélisations que nous avons vues se retrouveront dans les domaines de la recherche par le contenu dans des bases (Content Based Image Retrieval) ou l'appariement de formes (pattern matching). En particulier, la modélisation de la notion de similarité par projection dans un espace métrique est une démarche très répandue. Cependant, même si ces modélisations construites sur une distance permettent de faire le lien avec un arsenal mathématique très utile à l'analyse des données projetées dans ces espaces, il faut bien garder à l'esprit que la notion de similarité perceptive ne s'apprivoise pas si facilement. En effet, les différentes propriétés à l'origine de la définition d'une distance entre points d'un espace métrique ne sont pas vérifiées par la similarité perceptive entre stimuli en général, et entre images en particulier.

En particulier, il est commun de constater que certaines images sont plus auto-similaires que d'autres, contredisant donc la première propriété précédente. De même, on peut trouver des contre exemples à la propriété de symétrie ou d'inégalité triangulaire, comme nous le voyons sur la figure 5.1.

Figure 5.1. Exemple de non respect de l'inégalité triangulaire. L'image de gauche *(A)* et l'image de droite *(C)* sont jugées relativement dissemblables. En revanche, celle du milieu *(B)* est à la fois similaire à *(A)* et à *(C)*. La distance $d(A,C)$ serait donc supérieure à la somme $d(A,B) + d(B,C)$.

Un bon modèle de similarité devrait donc permettre de retrouver la possibilité de non vérification de ces propriétés mathématiques.

2.2 Les images : un stimulus particulier

Ces considérations physio-psychologiques ont eu une répercussion directe sur la manière dont la similarité a été pensée puis appliquée à des cas concrets au domaine de l'image. Cette notion de similarité est ainsi devenue une mesure calculable entre deux images portant généralement sur la construction de vecteurs résumant l'information visuelle contenue dans chaque image. Le système engagé dans la comparaison des images porte alors sur la mise au point d'une fonction de comparaison dans son espace de caractéristiques, qui joue le rôle de métrique ou de mesure de similarité. Cette fonction ne vérifie pas nécessairement les axiomes des distances, comme nous l'avons vu, en raison des fortes corrélations qu'elle a avec les propriétés du SVH (système visuel humain), qui ne vérifie pas ces propriétés *(Tversky, [1977]).* Nous avons en particulier noté que la propriété de symétrie n'est pas systématique lors de comparaison d'image par le SVH *(Scassellati et al. [1994]).* Il en est de même pour l'inégalité triangulaire qui ne rend pas compte de la réalité dans le cas de dissimilarités importantes entre images *(Hagedoorn et al. [2000]).*

3 Dictionnaires de formes et méthodes de comparaison

3.1 Représentation théorique des dictionnaires de formes

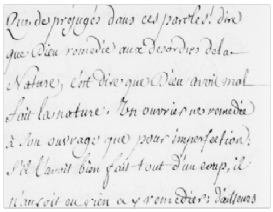
Étant donné un espace de représentation *cb* et des regroupements $C = C_1,\ldots, C_n$ de graphèmes $x_1,\ldots,x_k$ d'une page de texte manuscrit p_m, le dictionnaire de formes $cb(p_m)$ est défini comme l'ensemble des regroupements par :

$$cb(p_m) = \left\{\left(c_i^m, w_i^m\right), i = 1,\ldots,n\right\}, \text{ où } c_i^m = \frac{\sum_{x \in C_i} x}{|c_i|} \text{ et } w_i^m = \frac{|c_i|}{k} \qquad \textbf{(5.1)}$$

avec k le nombre total de graphèmes dans le dictionnaire de formes, représentent respectivement le centroïde et le poids.

Intuitivement, le dictionnaire de formes $cb(p_m)$ d'une page de texte manuscrit p_m représente l'ensemble des centroïdes c_i^m avec les poids w_i^m des clusters C_i. A titre d'exemple nous montrons la figure 5.2 qui présente un manuscrit p_m avec son dictionnaire de formes $cb(p_m)$ complet, composé de *15* regroupements.

(a) Manuscrit *m*



(b) $cb(p_m) = \{(c_i^m, w_i^m), i = 1,...,15\}, n = 15$

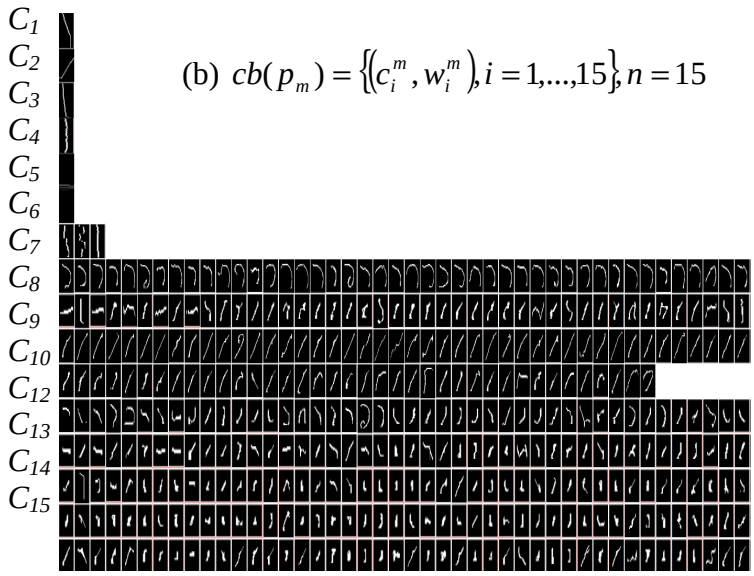


Figure 5.2. (a) Manuscit *m*, (b) représentation du dictionnaire de formes

Voyons à présent comment exploiter les différences entre deux dictionnaires de formes afin d'estimer leur similarité globale évaluée à l'aide de différentes distances que nous présentons dans les sections suivantes. Ces différentes distances présentent l'intérêt de permettre la comparaison d'ensembles de tailles différentes, ce qui est une condition essentielle pour comparer deux dictionnaires dont le nombre de classes et le nombre d'occurrences de graphèmes par classe sont variables.

3.2 Distance de Hausdorff

Étant donné deux dictionnaires de formes $cb(p_m)$ et $cb(p_k)$, et une distance de similarité d, la distance de Hausdorff HD est définie par :

$$HD(cb(p_m), cb(p_k)) = \max\{h(cb(p_m), cb(p_k)), h(cb(p_k), cb(p_m))\} \text{ où} \quad \textbf{(5.2)}$$
$$h(cb(p_m), cb(p_k)) = \max_{c^m \in cb(pm)} \min_{c^k \in cb(pk)} \{d(c^m, c^k)\}$$

Cette distance de Hausdorff est basée seulement sur le centroïde et ne prend pas compte les pondérations. Par ailleurs, elle ne tient pas compte de l'information concernant la totalité de la signature, ce qui cause une limitation (perte de contenu) lors du calcul de la distance entre deux signatures *(Beecks et al. [2010]).*

3.3 Distance de Hausdorff modifiée (Schaefer)

Étant donné deux dictionnaires de formes $cb(p_m)$ et $cb(p_k)$, la distance de Schaefer (ou distance de Hausforff modifiée MHD) est définie par:

$$MHD = \max\{h(cb(p_m), cb(p_k)), h(cb(p_k), cb(p_m))\} \text{ où}$$

$$h(cb(p_m), cb(p_k)) = \frac{1}{N}\sum_{i=1}^{N^m} \min_j \{d(c_i^m, c_i^k)\} \quad \textbf{(5.3)}$$

est la distance de Hausdorff modifiée de $cb(p_m)$ vers $cb(p_k)$. Il est possible d'exploiter d'autres distances d (que la distance euclidienne entre deux caractéristiques) comme nous le verrons dans les expérimentations qui suivront. Dans cette expression, nous voyons que, plutôt que de prendre le « maximum du minimum des distances entre dictionnaire de formes » comme c'est le cas dans la distance originale de Hausdorff, on utilise la moyenne des minimums ce qui rend la mesure de distance moins sensible aux valeurs aberrantes. Cette distance est similaire à la distance de Hausdorff, elle ne prend pas en compte les corrélations entre les données ni l'information globale accessible à partir de la totalité de la structure *(Schaefer, [2002])*.

3.4 Distance perpétuelle modifiée de Hausdorff

Étant donné deux dictionnaires de formes $cb(p_m)$ et $cb(p_k)$, et une distance de similarité d, la distance perpétuelle modifiée de Hausdorff *DPMH* est définie par :

$$DPMH(cb(pm), cb(pk)) = \max\{h(cb(p_m), cb(p_k)), h(cb(p_k), cb(p_m))\} où$$

$$h(cb(p_m), cb(p_k)) = \frac{\sum_i w_i^k . \min_j \left\{ \frac{d(c_i^k, c_j^m)}{\min(w_i^k, w_j^m)} \right\}}{\sum_i w_i^k} \quad \textbf{(5.4)}$$

Cette distance prend en considération le centroïde et le poids de chaque regroupement, mais tout comme la distance de Hausdorff, elle ne prend pas compte l'information de tout le dictionnaire de formes (corrélation entre les données) (Beecks et al. [2010]).

3.5 Earth Mover's Distance

Étant donné deux dictionnaires de formes $cb(p_m)$ et $cb(p_k)$, la distance *EMD (Levina et Bickel, [2001])* entre ces dictionnaires de formes est définie comme un flux de coût minimum sur tous les flux possibles f_{ij} est définie par :

$$EMD(cb(p_m), cb(p_k) = \min_{f_{ij}} \left\{ \frac{\sum_i \sum_j f_{ij} \cdot d\left(c_i^m, c_j^k\right)}{\min\left\{\sum_i w_i^m, \sum_j w_j^k\right\}} \right\} \quad \textbf{(5.5)}$$

Avec les contraintes :

$$\forall i : \sum_j f_{ij} \leq w_i^k, \forall j : \sum_j fij \leq w_j^m, \forall i, j : fi_j \geq 0 \text{ et } \sum_i \sum_j f_{ij} = \min\left\{\sum_i w_i^m, \sum_j w_j^k\right\}$$

Les contraintes garantissent une solution réalisable, assurant que tous les coûts soient positifs et ne dépassent pas les limites données par les poids dans les deux dictionnaires de formes. Cependant, le problème de minimisation à résoudre dans la formulation générale de la distance induit une complexité à l'exécution considérablement élevée *(Beecks et al. [2010]).*

3.6 Distance de corrélation pondérée

Étant donné deux dictionnaires de formes $cb(p_m)$ et $cb(p_k)$,un rayon maximal R et une distance d pour chaque regroupement, la distance de corrélation pondérée *DCP* est définie par :

$$DCP(cb(p_m), cb(p_k)) = 1 - \sum_i \sum_j s(c_i^k, c_j^m) \cdot \frac{w_i^k}{\sqrt{cb(p_k) \cdot cb(p_k)}} \cdot \frac{w_i^m}{\sqrt{cb(p_m) \cdot cb(p_m)}} \text{ où}$$

$$cb(p_k) \cdot cb(p_k) = \sum_i \sum_j s(c_i^k, c_j^m) \cdot w_i^k \cdot w_j^m \text{ et} \quad \textbf{(5.6)}$$

$$s(c_i, c_j) = \left\{ \begin{array}{l} 1 - \frac{3}{4}\frac{d}{R} + \frac{1}{16}\left(\frac{d}{R}\right)^3, 0 \leq \frac{d}{R} \leq 2 \\ 0, \text{sinon} \end{array} \right\}$$

Basée sur l'intersection $s(c_i, c_j)$ entre les deux centroïdes c_i et c_j, la pondération corrélée $cb(p_k) \cdot cb(p_m)$ entre les caractéristiques des graphèmes des deux dictionnaires de formes est normalisée et est utilisée pour déterminer la distance. Elle présente l'intérêt d'exploiter les pondérations de chaque classe et d'assurer par définition la prise en compte des corrélations entre les données *(Beecks et al. [2010]).*

3.7 Distance Quadratique

Étant donné deux dictionnaires de formes $cb(p_m)$ et $cb(p_k)$, la distance quadratique *DQ* est définie par :

$$DQ(cb(p_m), cb(p_k)) = \sqrt{(w_k | -w_m).A.(w_k | -w_m)^T} \quad \textbf{(5.7)}$$

A est la matrice de similarité qui permettra de pondérer les différences entre les deux dictionnaires de formes. $w_k = (w_1^k, ..., w_n^k), w_m = (w_1^m, ..., w_n^m)$ forment les vecteurs de poids calculés à partir de la formule présentée dans la section 3.1, et $(w_k | -w_m) = (w_1^k, ..., w_n^k, -w_1^m, ..., -w_n^m)$ exprime la concaténation de w_k et $-w_m$.

La matrice de similarité *A* qui est déterminée dynamiquement pour chaque comparaison de deux signatures reflète les similarités entre les centroïdes des regroupements des signatures et elle est présentée par :

Les entrées de *A* dépendent de l'ordre des centroïdes dans lequel ils apparaissent dans le dictionnaire de formes et sont calculés à partir d'une fonction de similarité, par exemple $a_{ij} = d(c_i^m, c_i^k)$, avec *d* est une fonction euclidienne. De plus, la distance quadratique calcule la distance entre $cb(p_m)$ et $cb(p_k)$ en considérant les poids et les positions de n'importe quel centroïde c_m et c_q. C'est la raison pour laquelle on dit que la distance quadratique tient compte des informations présentes sur l'ensemble des données. La distance de Hausdorff et ces variances (Schaefer et distance perpétuelle) produisent néanmoins des résultats satisfaisants, mais considèrent partiellement la pondération et la position des centroïdes *(Beecks et al. [2009]), (Beecks et al. [2010]).*

Nous présentons dans la section suivante les résultats de notre méthode de reconnaissance de style sur une application CBIR.

4 Application à la reconnaissance de styles, basée sur la technique CBIR

Dans cette section, nous allons proposer d'une part de montrer la pertinence des trois mesures de similarité décrites à la section 3 (distance de Hausdorff, distance de Schaefer et mesure quadratique) pour comparer les dictionnaires de formes et d'autre part d'établir un comparatif de la meilleure distance retenue avec deux approches de classification par styles d'écritures paléographiques portant sur des descripteurs globaux : une approche portant sur une signature globale par coefficients de curvelets et une approche portant sur la quantification des

informations de texture par les descripteurs du second ordre issus des matrices de cooccurrence d'Haralick. Les travaux servant de comparaison à notre approche ont été présentés *dans (Joutel et al. [2008])* et *(Siddiqi et Vincent, [2009])* et s'inscrivent dans le projet ANR Graphem qui soutient ce travail de doctorat.

4.1 Comparatifs des performances des trois principales mesures de similarités entre dictionnaires : DHD, DHDM et DQ

Dans cette application nous avons utilisé *310* images de la base de l'IRHT, et *140* images de manuscrits de la base d'Oxford. Comme distance de similarité entre les dictionnaires de formes nous avons comparé les distances de Hausdorff (*DHD*), Schaefer ou Hausdorff modifié (*DHDM*) et Quadratique (*DQ*) définies dans la section 3.

Les dictionnaires de formes ont été conçus à partir d'une pondération générique portant sur toute la base de test ce qui accélère considérablement le processus, plutôt que d'établir une pondération spécifique à chaque classe de styles nécessitant une étape d'apprentissage plus longue. Globalement les résultats portant sur la pondération globale générique sont similaires à ceux obtenus par une pondération par styles.

Pour tester cette représentation nous comparons une image requête d'un style donné à tous les manuscrits de la base à laquelle il appartient. La précision ***P*** et le rappel ***R*** des images correctement retournées sont obtenus directement à partir de la matrice de confusion des réponses pertinentes et non pertinentes retournées (comme illustrée au tableau 1). La *F*-mesure que nous calculons dans cette section combine la précision ***P*** et le rappel ***R***, elle est présentée par la formule suivante :

$$F_{measure} = 2 \times \frac{P \times R}{P + R} \qquad \textbf{(5.8)}$$

Top n : signifie les *n* premiers manuscrits de styles les plus proches du manuscrit requête. Les valeurs de précision et de rappel sont calculées sur ces *n* plus proches manuscrits (tableau 5.1).

Tableau 5.1. Résultats de précision et rappel pour les trois distances sur les bases de manuscrits d'Oxford et de l'IRHT

	Hausdorff		Schaefer		Quadratique	
	Précision	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**
	Base Oxford					
Top 5	*0,50*	*0,13*	*0,91*	*0,15*	*0,98*	*0,3*
Top 10	*0,40*	*0,21*	*0,90*	*0,25*	*0,96*	*0,49*
Top 15	*0,39*	*0,30*	*0,83*	*0,73*	*0,89*	*0,68*
Top 20	*0,37*	*0,37*	*0,77*	*0,85*	*0,78*	*0,80*

Les résultats sur la base d'Oxford montrent qu'en moyenne on a une bonne précision de récupération essentiellement avec la distance Quadratique qui a fourni les meilleurs résultats. En effet, pour le *Top 20* nous avons obtenu une précision de *78%* avec un rappel de *80%*, ce qui signifie que *80%* des manuscrits de même style que l'image requête ont été donnés par l'application CBIR avec une F-Mesure de *79%*. La distance de Schaefer fournit des résultats comparables avec pour le *Top 15* une précision de *83%* et un rappel de *73%* produisant une F-mesure de *77.6%*. En revanche, la distance de Hausdorff n'a pas donné de résultats satisfaisants. Cela montre la nécessité de prendre en compte les informations de pondérations des clusters (proportionnelles au nombre de représentants par classes) ainsi que les données centroïdes des classes pour la comparaison des dictionnaires de formes.

La figure 5.3 montre les courbes de Précision-Rappel pour la base d'Oxford. A partir de ces courbes nous remarquons l'avantage des résultats obtenus à partir de la distance Quadratique.

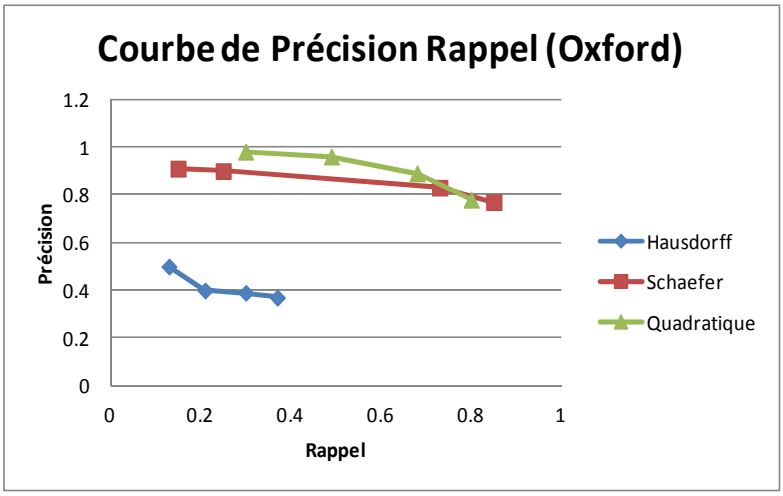


Figure 5.3. Courbes de précision-rappel, base Oxford

Dans la section suivante nous comparons notre méthode hybrique (locale et globale) en utilisant les trois distances de similarités avec les autres méthodes utilisées dans le projet Graphem dont la méthode globale basée sur les curvelets de *(Joutel et al. [2008])* et la méthode hybrides basée aussi sur les dictionnaires de formes et les chaines de freeman de *(Siddiqi et Vincent, [2009]).*

4.2 Analyse des performances de notre contribution

Afin de se positionner par rapport aux autres méthodes développées dans le cadre du projet Graphem et déduire si notre contribution produit desrésultats satisfaisants et exploitables sur les manuscrits médiévaux de la base IRHT, nous avons choisi de réaliser différents tests comparatifs. Dans un premier temps, nous avons produit un comparatif des performances de notre approche en exploitant tour à tour les différentes métriques introduites pour la comparaison des dictionnaires de formes issus de notre méthodologie (distances de Hausdorff, de Scaefer et mesure quadratique).

Dans un second temps, nous avons produit un tableau de comparaison des performances en taux de Précision et de Rappel par rapport aux approches concurrentes du projet : l'approche par analyse par Curvelets *(Joutel et al. [2008])* et par l'approche par analyse de contours *(Siddiqi et Vincent, [2009])*. Nous avons ainsi estimé les taux de Précision et Rappel pour les Top-N images les plus similaires à une image requête, en suivant le même principe que celui proposé dans la section précédente.

Les résulats de Précision et Rappel sont présentés dans le tableau 5.2. Les trois premières colonnes représentent nos résultats utilisant la distance de Hausdorff, Schaefer et Quadratique. Les deux dernières colonnes présentent respectivement les résultats de la méthode hybride proposée par *(Siddiqi et Vincent, [2009])* et la méthode de curvelets globale proposée par *(Joutel et al. [2008]).*

Tableau 5.2. Résultats de précision et rappel pour les trois distances sur les bases de manuscrits d'Oxford et de l'IRHT

	Hausdorff		**Schaefer**		**Quadratique**		**(Siddiqi et Vicent, [2009])**		**(Joutel et al. [2008])**	
	Précision	***Rappel***	***Précision***	***Rappel***	***Précision***	***Rappel***	***Précision***	***Rappel***	***Precision***	***Rappel***
	Base IRHT									
Top *5*	*0,3*	*0,05*	*0,37*	*0,04*	*0,43*	*0,065*	*0,6*	*0,22*	*0,29*	*0,03*
Top *10*	*0,2*	*0,06*	*0,35*	*0,06*	*0,33*	*0,09*	*0,5*	*0,35*	*0,29*	*0,07*
Top *15*	*0,18*	*0,07*	*0,33*	*0,09*	*0,29*	*0,12*	*0,44*	*0,48*	*0,30*	*0,10*
Top *20*	*0,17*	*0,1*	*0,22*	*0,18*	*0,26*	*0,15*	*0,42*	*0,55*	*0,29*	*0,15*

Nous constatons que les résultats produits sur la base de l'IRHT n'ont pas été à la hauteur de nos attentes. La Méthode de *(Siddiqi et Vicent, [2009])* est celle qui fournit les meilleurs taux de Précision et de Rappel, mais demeurent globalement insatisfaisants en ne révélant pas toute la pertinence de l'approche proposée. La méthode proposée par *(Joutel et al. [2008])* comme notre méthode produit des résultats très faibles sur cette base médiévale et en dessous des

résultats produits par Siddidi dans *(Siddiqi et Vicent, [2009]).* Ces faibles résultats sont néanmoins prévisibles, en raison de l'absence d'une vérité terrain visuelle mais exclusivement construite à partir du classement réalisé par les spécialistes paléographes et conduisant à l'établissement de frontières instables entre classes : il est en effet possible de trouver des manuscrits qui appartiennent à deux classes d'écritures différentes, ce qui impacte à la baisse le taux de classification.

La méthode globale de *(Joutel et al. [2008])* basée sur informations d'orientation et de courbure calculée à partir des curvelets, ne fournit pas suffisamment d'informations pour pouvoir avoir un grand pouvoir de discrimination, spécialement quand il existe des frontières instables entre les classes ou lorsque certaines orientations trop faiblement représentées deviennent insuffisamment discriminantes.

Notre méthode hybride comme la celle de *(Siddiqi et Vicent, [2009])*, est basée en partie sur les dictionnaires de formes. Dans notre cas, nous n'avons pas exploité l'information de contours comme c'est le cas dans les travaux de Siddiqi et Vincent. Il faut noter que les auteurs ont introduit un ordonnancement dans le suivi de tracé du contour (à travers le codage de Freeman) ce qui de toute évidence est un facteur influençant la bonne discrimination des classes d'écritures. La figure 5.18 montre les courbes de Précision-Rappel pour la base de l'IRHT.

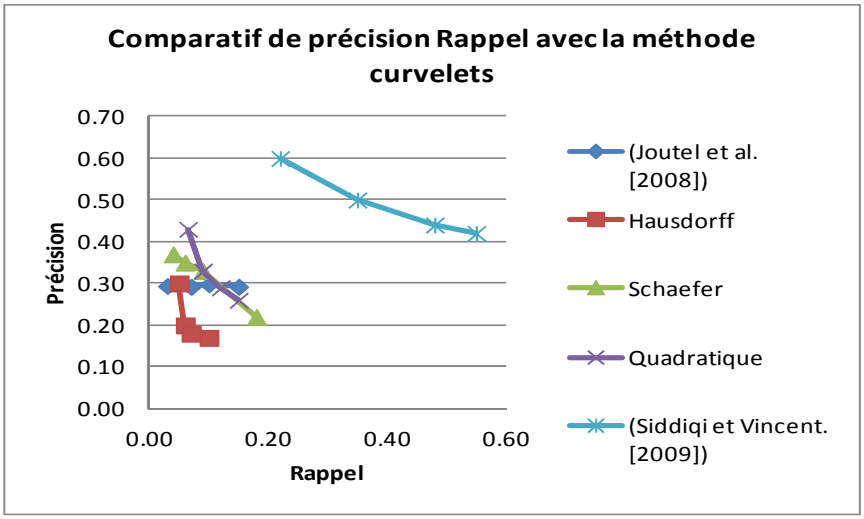


Figure 5.4. Courbes de précision-rappel, base IRHT

Sur la figure 5.18, la méthode de Siddiqi et Vincent produit les meilleurs résultats. La nature des descripteurs retenus pour cette base est incontestablement essentielle pour produire ces résultats. Voyons à présent comment se positionne notre approche dans un contexte étendu à une autre base étiquettée visuellement et à un plus grand panel d'écritures.

5 Comparaison des résultats portant sur les poids génériques et les dictionnaires de formes représentatifs de chaque style

Dans le chapitre précédent, nous avons montré notre méthode de recherche de dictionnaires de formes représentatifs pour chaque style, et avons calculé la distance entre les poids génériques spécifiques à chaque classe. Dans cette section nous montrons d'une part les résultats des précisions à partir d'une application CBIR en utilisant les poids génériques et les dictionnaires de formes représentatifs de chaque classe et d'autre part les résultats de précisions utilisant les poids génériques sur toute la base et sans dictionnaire de forme représentatif. La précision est calculée de la même façon que la section précédente, et en utilisant les distances DHD, DHDM et DQ.

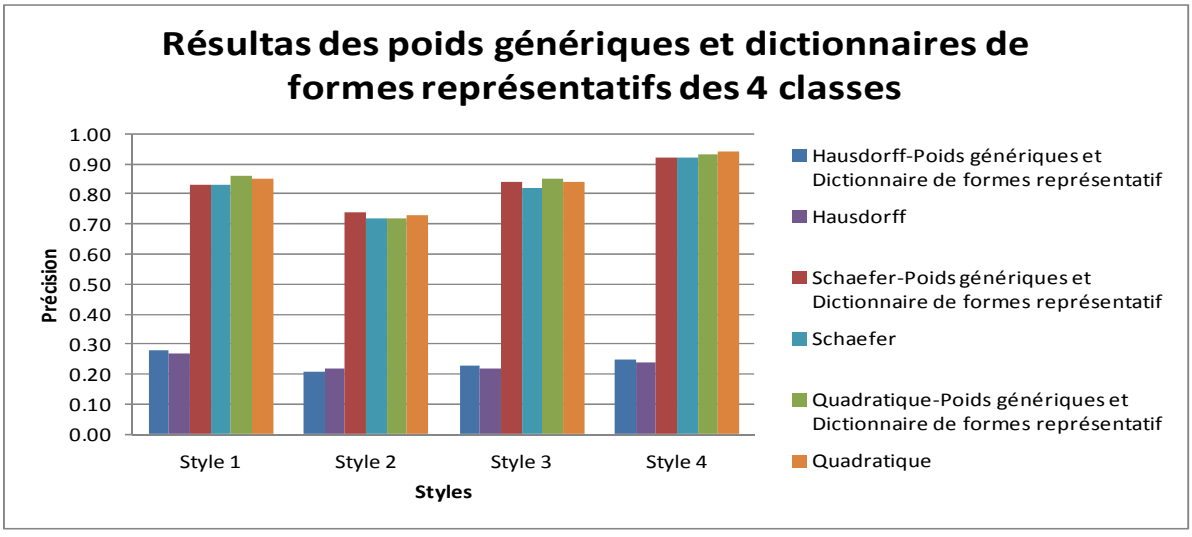


Figure 5.5. Résultats des poids génériques et dictionnaires de formes spécifiques à chaque style et poids génériques sur toute la base sans dictionnaires de formes représentatifs

La figure 5.5 montre des précisions de CBIR par rapport à chacun des 4 styles de la base d'Oxford, en utilisant les dictionnaires de formes et poids génériques relatifs au style (obtenus sur chaque classe de style) et les dictionnaires de formes et poids génériques relatifs à toute la base (obtenus sur toute la base). On voit que les deux approches offrent des précisions

semblables. Cela montre que sur une base dimensionnée comme la base d'Oxford comportant quelques dizaines de styles d'écritures maximum, il n'est pas nécessaire de repartitionner à priori la base selon les différents styles (cette tâche fastidieuse est souvent effectuée manuellement par des paléographes) pour obtenir des poids génériques offrant les mêmes performances.

6 Positionnement de notre méthode : compétition ICDAR 2011

Dans cette section nous allons comparer notre méthode d'identification de scripteurs aux huit méthodes participant à la compétition d'ICDAR 2011, en utilisant les mêmes métriques d'évaluation et la même base d'images. Cette comparaison nous permet de nous positionner par rapport à ces méthodes récentes qui utilisent différentes signatures globales ou locales et distances entre manuscrits. Parmi ces méthodes une seule est basée sur un dictionnaire de formes sous le nom de CS-UMD. Notons que la méthode NCS-NUST est proposée par Siddiqi mais elle utilise seulement les caractéristiques locales basées sur les contours sans les dictionnaires de formes. Nous présentons au début de cette section un bref descriptif de chacune de ces méthodes.

6.1 Description des méthodes

Commençons tout d'abord par un résumé sur les méthodes utilisées dans cette compétition d'identification de scripteur.

1. **Méthode ECNU :** Cette méthode utilise les caractéristiques basées sur la direction des contours qui encode l'orientation et la courbure dans une grille locale autour de chaque pixel du contour. La métrique du khi-deux est utilisée pour mesurer la distance entre manuscrits.
2. **Méthode QUQA-a :** Cette méthode utilise d'une part la distribution de probabilités des caractéristiques directionnelles basées sur le contour et d'autre part des caractéristiques basées sur les graphèmes. L'étape de classification est effectuée en utilisant un classifieur de régression logistique qui est appliqué directement sur tout le document.
3. **Méthode QUQA-b :** Cette méthode utilise les mêmes caractéristiques que la méthode QUQA-a, mais le classifieur de régression logistique est appliqué cette fois sur le document après décomposition en quatre blocs.
4. **Méthode TSINGHUA :** Cette méthode utilise une grille de micro-caractéristiques et analyse les textes manuscrits en multi-ligne. Un ensemble de micro-caractéristiques

est calculé en utilisant une fenêtre de grille mobile. La variance de la distance du chi-deux pondérée est utilisée pour l'identification des scripteurs.

5. **Méthode GWU :** Cette méthode combine neuf caractéristiques : code chaine, direction, épaisseur… Pour comparer deux manuscrits, les auteurs utilisent la métrique de Mahalanobis modifiée.

6. **Méthode CS-UMD :** Les caractéristiques K-adjacent de segment sont utilisées dans un sac de caractéristiques (SDC) afin de modéliser le manuscrit. Le modèle SAC est utilisé pour comparer les scripteurs de deux manuscrits en convertissant les caractéristiques extraites d'un document en un histogramme de codes de mots. L'histogramme est normalisé pour qu'il soit invariant à la taille des entrées. Les histogrammes sont comparés deux à deux en utilisant la distance euclidienne.

7. **Méthode TEBESSA :** Cette méthode utilise la distribution de probabilités des run-lengths en noir et blanc dans quatre directions (horizontale, verticale, diagonale gauche et droite). L'histogramme des run-lengths est normalisé et interprété comme une distribution de probabilités. Pour comparer deux manuscrits, ils utilisent la distance de Manhattan.

8. **Méthode NCS-NUST :** Cette méthode est basée sur un ensemble de caractéristiques qui capturent l'information d'orientation et de courbure sur différents niveaux d'observation. Ces caractéristiques sont calculées à partir de contours de l'image représentée par les codes de Freeman ainsi que par un ensemble de polygones. Au total *14* caractéristiques sont extraites, notamment l'histogramme des chaines de Freeman avec leurs différentielles, l'index de la courbure, les histogrammes pondérés et non pondérées d'inclinaison et de courbure de segments de ligne approximant les contours. La distance entre deux documents est définie comme la moyenne de toutes les distances entre les caractéristiques.

6.2 Notre contribution en deux méthodes

Pour le positionnement de notre méthode par rapport aux différentes techniques évaluées dans le cadre de la compétition ICDAR 2011, nous avons effectué deux grandes expérimentations de comparaison des dictionnaires de formes en exploitant deux distances de similarité qui ont fourni les meilleurs résultats dans les tests précédents. Pour cela nous avons prévu deux approches (nommées méthode 1 et méthode 2) :

1. **La Distance Quadratique (méthode 1):** Chaque manuscrit est décomposé en graphèmes en suivant les règles de décomposition présentées dans le chapitre 2 (minimum global, croisement, trait déjà visité). Puis chaque graphème est caractérisé

par un vecteur de *59* caractéristiques. Les graphèmes sont ensuite classifiés de façon non-supervisée en groupes homogènes pour former un dictionnaire de formes. La classification est effectuée par un classifieur basé sur l'algorithme de coloration de graphe. La distance entre deux dictionnaires de formes est calculée à partir de la distance quadratique en tenant compte du centroïde de chaque classe et des pondérations.

2. **La Distance de Schaefer (méthode 2):** nous suivons ici le même principe que la méthode précédente pour la construction des dictionnaires de formes. La distance entre deux dictionnaires est calculée en utilisant la distance de Schaefer qui prend en compte le centroïde sans les pondérations des classes.

6.3 Description de la base ICDAR 2011

La base d'images de test de la compétition ICDAR 2011 a été produite à partir de *26* scripteurs à qui il a été demandé de copier huit pages de texte en plusieurs langues (anglais, français, allemand, grecque). Toutes les images ont été binarisées et ne contiennent pas d'éléments non textuels (lignes, dessins, etc.) (figure 5.6).

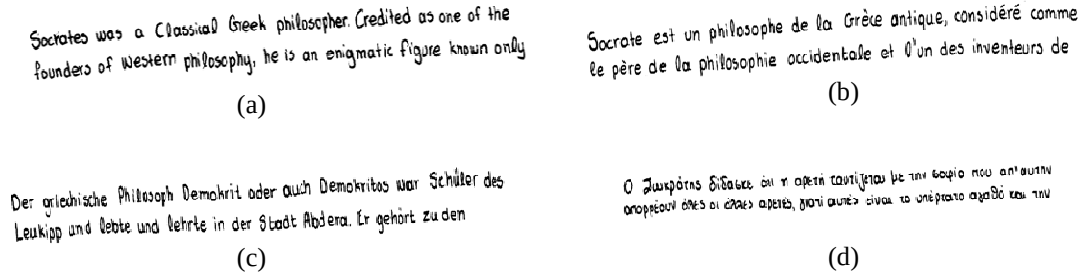


Figure **5.**6. Quatre échantillons de la base ICDAR 2011 écrits en : (a) anglais, (b) français, (c) allemand, (d) grec

6.4 Métrique d'évaluation

Afin de mesurer la précision des méthodes présentées, nous utilisons les critères TOP-N souple et TOP-N strict.

Pour le critère TOP-N souple, on considère qu'on a une bonne classification si *au moins* une image de manuscrit du même scripteur est comprise dans les *N* images de manuscrit les plus similaires.

Pour le critère TOP-N strict, on considère qu'on a une bonne classification si *toutes* les *N* images de manuscrits les plus similaires sont effectivement écrites par le même scripteur.

Pour toutes les images de manuscrits présentes dans la base, on compte ensuite le nombre de bonnes classifications. Le *quotient* du nombre total de bonnes classifications sur le nombre total d'images dans la base correspond à la précision TOP-N.

Les valeurs de N utilisées pour le critère souple sont : *1, 2, 5, 10*. Comme la base est composée de 8 images par scripteurs, la valeur *7* correspond à la valeur maximum de *N* pour le critère strict.

Pour chaque critère (souple ou strict), le rang de chaque méthode est calculé. Le rang final est calculé après le tri de la valeur accumulée de classement de tous les critères. Plus spécifiquement, soit *R(j)* le rang obtenu pour une méthode pour le critère *j*. où *j=1…m*, avec *m* le nombre total de critères. Pour chaque méthode d'identification de scripteur, le rang final *S* est calculé en additionnant les *m* rangs. La meilleure méthode va avoir la plus petite valeur de *S*.

$$S = \sum_{j=1}^{m} R(j) \qquad \textbf{(5.9)}$$

6.5 Évaluation des méthodes

L'évaluation des méthodes est basée sur deux scénarios.

1/ Dans le premier scénario l'image du manuscrit toute entière est utilisée. Les résultats d'évaluation de toutes les méthodes portant sur la base toute entière sont présentés dans les tableaux 5.3 et 5.4.Les résultats d'évaluation pour chaque langue prise indépendamment des autres sont présentés dans les tableaux 5.5-5.8. Dans toutes les tables, les méthodes qui ont la plus grande précision sont marquées en gras, et le classement de chaque méthode est présenté entre parenthèses. Nos deux propositions de méthodes (méthode 1 et 2) sont marquées en gris. Concernant le premier scénario, la méthode de *TSINGHUA* a donné les meilleurs résultats avec une valeur minimale de *S* correspondant à une valeur de *m = 23*.

Tableau 5.3. Evaluation souple en utilisant toute la base (%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	*98,1(4)*	*98,6(3)*	***100(1)***	***100(1)***
méthode 2	*98,1(4)*	*98,6(3)*	***100(1)***	***100(1)***
ECNU	*84,6(7)*	*86,5(6)*	*88,0(4)*	*88,9(4)*
QUQA-a	*90,9(6)*	*94,2(5)*	*98,1(3)*	*99,0(3)*
QUQA-b	*98,1(4)*	*98,6(3)*	*99,5(2)*	***100,0(1)***
TSINGHUA	***99,5(1)***	*99,5(2)*	***100,0(1)***	***100,0(1)***
GWU	*93,8(5)*	*96,2(4)*	*98,1(3)*	*99,0(3)*
CS-UMD	***99,5(1)***	*99,5(2)*	*99,5(2)*	*99,5(2)*
TEBESSA	*98,6(3)*	***100,0(1)***	***100,0(1)***	***100,0(1)***
MCS-NUST	*99,0(2)*	*99,5(2)*	*99,5(2)*	*99,5(2)*

Tableau 5.4. Evaluation stricte en utilisant toute la base (%)

Méthode	*TOP-2*	*TOP-5*	*TOP-7*
méthode 1	*85,1(6)*	*43,3(7)*	*19,2(6)*
méthode 2	*80,8(7)*	*38,5(9)*	*8,2(7)*
ECNU	*51,0(10)*	*2,9(10)*	*0,0(8)*
QUQA-a	*76,4(9)*	*42,3(8)*	*20,2(5)*
QUQA-b	*92,3(4)*	*77,4(5)*	*41,4(2)*
TSINGHUA	*95,2(2)*	***84,1(1)***	*41,4(2)*
GWU	*80,3(8)*	*44,2(6)*	*20,2(5)*
CS-UMD	*91,8(5)*	*77,9(4)*	*22,1(4)*
TEBESSA	***97,1(1)***	*81,3(2)*	***50,0(1)***
MCS-NUST	*93,3(3)*	*78,9(3)*	*38,9(3)*

Tableau 5.5. Evaluation souple sur les manuscrits grecs (%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	***100(1)***	***100(1)***	***100(1)***	***100(1)***
méthode 2	***100(1)***	***100(1)***	***100(1)***	***100(1)***
ECNU	*19,2(9)*	*19,2(7)*	*19,2(6)*	*21,2(5)*
QUQA-a	*76,9(8)*	*86,5(6)*	*96,2(3)*	*98,1(2)*
QUQA-b	*90,4(5)*	*90,4(4)*	*92,3(4)*	*94,2(4)*
TSINGHUA	*92,3(4)*	*94,2(3)*	*98,1(2)*	***100,0(1)***
GWU	*80,8(7)*	*86,5(6)*	*90,4(5)*	*94,2(4)*
CS-UMD	*96,2(2)*	*96,2(2)*	*96,2(3)*	*96,2(3)*
TEBESSA	*84,6(6)*	*88,5(5)*	*90,4(5)*	*94,2(4)*
MCS-NUST	*94,2(3)*	*94,2(3)*	*96,2(3)*	*96,2(3)*

Tableau 5.6. Evaluation souple sur les manuscrits anglais(%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	***100(1)***	***100(1)***	***100(1)***	***100(1)***
méthode 2	***100(1)***	***100(1)***	***100(1)***	***100(1)***
ECNU	*15,4(6)*	*15,4(5)*	*15,4(4)*	*17,3(4)*
QUQA-a	*78,9(5)*	*84,6(4)*	*96,2(3)*	*96,2(3)*
QUQA-b	***100,0(1)***	***100,0(1)***	***100,0(1)***	***100,0(1)***
TSINGHUA	*96,2(3)*	*96,2(2)*	*98,1(2)*	***100,0(1)***
GWU	*84,6(4)*	*88,5(3)*	*96,2(3)*	*98,1(2)*
CS-UMD	*98,1(2)*	***100,0(1)***	***100,0(1)***	***100,0(1)***
TEBESSA	*96,2(3)*	*96,2(2)*	*98,1(2)*	***100,0(1)***
MCS-NUST	*96,2(3)*	*96,2(2)*	*98,1(2)*	***100,0(1)***

Tableau 5.7. Evaluation souple sur les manuscrits français(%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	*98,1(2)*	***100(1)***	***100(1)***	***100(1)***
méthode 2	*98,1(2)*	***100(1)***	***100(1)***	***100(1)***
ECNU	*23,1(6)*	*23,1(5)*	*23,1(3)*	*26,9(2)*
QUQA-a	*94,2(4)*	*96,2(3)*	*96,2(2)*	***100,0(1)***
QUQA-b	*98,1(2)*	*98,1(2)*	***100,0(1)***	***100,0(1)***
TSINGHUA	*96,2(3)*	*98,1(2)*	***100,0(1)***	***100,0(1)***
GWU	*96,2(3)*	*96,2(3)*	***100,0(1)***	***100,0(1)***
CS-UMD	***100,0(1)***	***100,0(1)***	***100,0(1)***	***100,0(1)***
TEBESSA	*92,3(5)*	*94,2(4)*	***100,0(1)***	***100,0(1)***
MCS-NUST	***100,0(1)***	***100,0(1)***	***100,0(1)***	***100,0(1)***

Tableau 5.8. Evaluation souple sur les manuscrits allemands(%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	*98,1(2)*	***100(1)***	***100(1)***	***100(1)***
méthode 2	*98,1(2)*	***100(1)***	***100(1)***	***100(1)***
ECNU	*46,2(6)*	*46,2(5)*	*46,2(3)*	*46,2(2)*
QUQA-a	*86,5(5)*	*90,4(4)*	*98,1(2)*	***100,0(1)***
QUQA-b	***100,0(1)***	***100,0(1)***	***100,0(1)***	***100,0(1)***
TSINGHUA	***100,0(1)***	***100,0(1)***	***100,0(1)***	***100,0(1)***
GWU	*92,3(4)*	*94,2(3)*	*98,1(2)*	***100,0(1)***
CS-UMD	***100,0(1)***	***100,0(1)***	***100,0(1)***	***100,0(1)***
TEBESSA	*94,2(3)*	*98,1(2)*	***100,0(1)***	***100,0(1)***
MCS-NUST	***100,0(1)***	***100,0(1)***	***100,0(1)***	***100,0(1)***

2/ Dans le deuxième scénario, les images des manuscrits sont recadrées en préservant seulement les deux premières lignes, afin de diminuer la quantité d'informations disponibles. Puis, les mêmes expériences que dans le premier scénario sont répétées mais cette fois avec les images recadrées. Dans le deuxième scénario aussi la méthode de *TSINGHUA* a donné les meilleurs résultats.

Tableau 5.9. Evaluation souple en utilisant toute la base des images recadrées (%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	***100(1)***	***100(1)***	***100(1)***	***100(1)***
méthode 2	***100(1)***	***100(1)***	***100(1)***	***100(1)***
ECNU	*65,9(8)*	*71,6(8)*	*81,7(8)*	*86,5(8)*
QUQA-a	*74,0(5)*	*81,7(5)*	*91,8(5)*	*96,2(4)*
QUQA-b	*67,3(6)*	*79,8(6)*	*91,8(5)*	*94,7(6)*
TSINGHUA	*90,9(2)*	*93,8(2)*	*98,6(2)*	*99,5(2)*
GWU	*74,0(5)*	*81,7(5)*	*91,4(6)*	*95,2(5)*
CS-UMD	*66,8(7)*	*75,5(7)*	*83,7(7)*	*89,9(7)*
TEBESSA	*87,5(3)*	*92,8(3)*	*97,6(3)*	*99,5(2)*
MCS-NUST	*82,2(4)*	*91,8(4)*	*96,6(4)*	*97,6(3)*

Tableau 5.10. Evaluation stricte en utilisant toute la base des images recadrées(%)

Méthode	*TOP-2*	*TOP-5*	*TOP-7*
méthode 1	*46,6(8)*	*7,2(8)*	*0,96(7)*
méthode 2	*35,6(10)*	*4,3(9)*	*1,9(6)*
ECNU	*39,4(9)*	*2,9(10)*	*0,0(8)*
QUQA-a	*52,4(4)*	*15,9(7)*	*3,4(5)*
QUQA-b	*47,6(7)*	*22,6(4)*	*6,3(4)*
TSINGHUA	***79,8(1)***	***48,6(1)***	*12,5(2)*
GWU	*51,4(6)*	*20,2(6)*	*6,3(4)*
CS-UMD	*51,9(5)*	*22,1(5)*	*3,4(5)*
TEBESSA	*76,0(2)*	*34,1(3)*	***14,4(1)***
MCS-NUST	*71,6(3)*	*35,6(2)*	*11,1(3)*

Tableau 5.11. Evaluation souple sur les manuscrits grecs recadrés (%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	***100(1)***	***100(1)***	***100(1)***	***100(1)***
méthode 2	***100(1)***	***100(1)***	***100(1)***	***100(1)***
ECNU	*11,5(8)*	*15,4(9)*	*19,2(9)*	*23,1(8)*
QUQA-a	*44,2(4)*	*51,9(6)*	*73,1(6)*	*90,4(5)*
QUQA-b	*34,6(7)*	*55,8(5)*	*76,9(5)*	*80,8(6)*
TSINGHUA	*51,9(3)*	*71,2(2)*	*98,1(2)*	*98,1(2)*
GWU	*42,3(5)*	*46,2(7)*	*65,4(8)*	*76,9(7)*
CS-UMD	*40,4(6)*	*44,2(8)*	*67,3(7)*	*76,9(7)*
TEBESSA	*42,3(5)*	*63,5(4)*	*80,8(4)*	*92,3(4)*
MCS-NUST	*55,8(2)*	*69,2(3)*	*84,6(3)*	*94,2(3)*

Tableau 5.12. Evaluation souple sur les manuscrits anglais recadrés (%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	***100(1)***	***100(1)***	***100(1)***	***100(1)***
méthode 2	***100(1)***	***100(1)***	***100(1)***	***100(1)***
ECNU	*13,5(9)*	*15,4(9)*	*15,4(7)*	*19,2(5)*
QUQA-a	*55,8(6)*	*67,3(6)*	*75,0(5)*	*82,7(4)*
QUQA-b	*63,5(5)*	*69,2(5)*	*90,4(3)*	*96,2(2)*
TSINGHUA	*76,9(2)*	*90,4(2)*	*96,2(2)*	***100,0(1)***
GWU	*50,0(7)*	*57,7(7)*	*69,2(6)*	*82,7(4)*
CS-UMD	*44,2(8)*	*50,0(8)*	*69,2(6)*	*82,7(4)*
TEBESSA	*69,2(3)*	*84,6(3)*	*88,5(4)*	***100,0(1)***
MCS-NUST	*67,3(4)*	*80,8(4)*	*88,5(4)*	*92,3(3)*

Tableau 5.13. Evaluation souple sur seulement les manuscrits français recadrés (%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	***100(1)***	***100(1)***	***100(1)***	***100(1)***
méthode 2	***100(1)***	***100(1)***	***100(1)***	***100(1)***
ECNU	*46,2(9)*	*46,2(7)*	*46,2(8)*	*46,2(7)*
QUQA-a	*51,9(7)*	*67,3(6)*	*88,5(5)*	*92,3(4)*
QUQA-b	*48,1(8)*	*67,3(6)*	*84,6(7)*	*88,5(6)*
TSINGHUA	*80,8(2)*	*90,4(2)*	*96,2(2)*	*96,2(2)*
GWU	*57,7(6)*	*69,2(5)*	*86,5(6)*	*92,3(4)*
CS-UMD	*59,6(5)*	*67,3(6)*	*84,6(7)*	*90,4(5)*
TEBESSA	*63,5(4)*	*78,9(4)*	*90,4(4)*	*94,2(3)*
MCS-NUST	*65,4(3)*	*82,7(3)*	*92,3(3)*	*96,2(2)*

Tableau 5.14. Evaluation souple sur les manuscrits allemands recadrés(%)

Méthode	*TOP-1*	*TOP-2*	*TOP-5*	*TOP-10*
méthode 1	***100(1)***	***100(1)***	***100(1)***	***100(1)***
méthode 2	***100(1)***	***100(1)***	***100(1)***	***100(1)***
ECNU	*21,2(8)*	*21,2(9)*	*23,1(8)*	*26,9(6)*
QUQA-a	*71,2(5)*	*78,9(5)*	*86,5(6)*	*94,2(4)*
QUQA-b	*44,2(7)*	*63,5(8)*	*84,6(7)*	*92,3(5)*
TSINGHUA	*84,6(2)*	*90,4(2)*	*96,2(2)*	***100,0(1)***
GWU	*69,2(6)*	*76,9(6)*	*88,5(5)*	*92,3(5)*
CS-UMD	*73,1(4)*	*80,8(4)*	*90,4(4)*	*96,2(3)*
TEBESSA	*71,2(5)*	*75,0(7)*	*88,5(5)*	*98,1(2)*
MCS-NUST	*78,9(3)*	*88,5(3)*	*94,2(3)*	*98,1(2)*

Le tableau suivant présente le rang de toutes les méthodes en termes de score S trié par ordre croissant sur les deux scénarios.

Tableau 5.15. Classement général en fonction des scores de *S* pour toutes les expériences. Les colonnes *2* à *13* reprennent les résultats des tableaux *2* à *13* précédents

Méthode	*5.3*	*5.4*	*5.5*	*5.6*	*5.7*	*5.8*	**S [Scénario *1*]**	*5.9*	*5.10*	*5.11*	*5.12*	*5.13*	*5.14*	**S [Scénario *2*]**	**S [Scénarios 1&2]**	**Rang**
TSINGHUA	5	5	10	8	7	4	**39**	8	4	9	7	8	7	**43**	**82**	**1**
méthode 1	9	19	4	4	5	5	**46**	4	23	4	4	4	4	**43**	**89**	**2**
méthode 2	9	23	4	4	5	5	**50**	4	25	4	4	4	4	**45**	**95**	**3**
MCS-NUST	8	9	12	8	4	4	**45**	15	8	11	15	11	11	**71**	**116**	**4**
TEBESSA	6	4	20	8	11	7	**56**	11	6	17	11	15	19	**79**	**135**	**5**
CS-UMD	7	13	10	5	4	4	**43**	28	15	28	26	23	12	**132**	**175**	**6**
QUQA-b	10	11	17	4	6	4	**52**	22	15	23	15	27	27	**129**	**181**	**7**
QUQA-a	17	22	19	15	10	12	**95**	19	16	21	21	22	20	**119**	**214**	**8**
GWU	15	19	22	12	8	10	**86**	21	16	27	24	21	22	**131**	**217**	**9**
ECNU	21	28	27	19	16	16	**127**	32	27	34	30	31	31	**185**	**312**	**10**

Nos méthodes ont été classées au bilan final comme deuxième et troisième juste derrière la méthode TSINGHUA. Dans le scénario 2, nous remarquons que malgré la diminution de performance de toutes les méthodes, nos deux méthodes ont pu garantir la présence d'au moins un manuscrit dans les TOP-N souple.

La distance DQ portant sur la forme quadratique a donné de meilleurs résultats que la distance de Schaefer, ce qui confirme nos premières expérimentations sur les bases paléographiques et l'importance des pondérations de chaque classe de graphèmes dans le calcul des distances entre dictionnaires de formes.

7 Conclusion

A travers ce chapitre nous clôturons notre travail d'étude des styles d'écritures.

Nous avons montré que notre méthode basée sur les dictionnaires de formes offrait de bons taux de classification de manuscrits par style évalués à partir d'applications CBIR, et cela malgré des résultats de classifications très faibles obtenus sur la base d'images paléographiques de l'IRHT, dus à la nature incertaine de la vérité terrain portant sur une dimension non exclusivement visuelle (le sens dégagé du texte, la présence de certains mots, l'organisation spatiale des lignes d'écriture et d'autres critères plus subjectifs ou sémantiques devraient à ce stade pouvoir être pris en considération pour permettre de contrôler une parfaite identification des styles).

Sur la base d'Oxford, les taux de classification ont été évalué à partir de l'estimation des poids génériques pris globalement sur l'ensemble des images de la base d'apprentissage (et non spécifiques à chaque style) ; l'objectif étant d'évaluer l'impact d'un calcul de poids spécifique à chaque style pouvant éventuellement améliorer les résultats de classement. Peu de changements ont pu être constatés. D'autres expérimentations devront encore être menées sur des bases d'images plus vastes présentant un nombre de classes d'écriture plus importants (quelques centaines)

A la fin de ce chapitre, nous avons finalement montré que notre approche, au-delà de sa conception dédiée aux écritures paléographiques, est en faite très générique : elle se classe respectivement en $2^{ème}$ et $3^{ème}$ place pour la distance Quadratique et la distance de Hausdorff modifiée sur les données de la compétition ICDAR 2011. Nous avons montré que la distance quadratique basée sur le centroïde et les pondérations fournit de meilleurs résultats que celle de Haussdorf spécialement lors de l'application des règles strictes d'identification de scripteurs (toutes les images de manuscrit les plus similaires à l'image requête doivent effectivement être récupérées dans les N premiers résultats). Cette dernière étude montre la faisabilité de notre méthode pour des applications ouvertes d'identification de scripteurs contemporains. Il s'agit d'une perspective très intéressante qui montre qu'au-delà des règles de décomposition des traces écrites, nos descripteurs et notre méthode de clustering basé sur la coloration de graphe sont des outils efficaces d'aide à la décision. Malgré les faiblesses rencontrées par notre proposition lorsque le nombre de lignes de texte est insuffisant, notre méthode se classe parmi les meilleures du domaine.

Conclusion

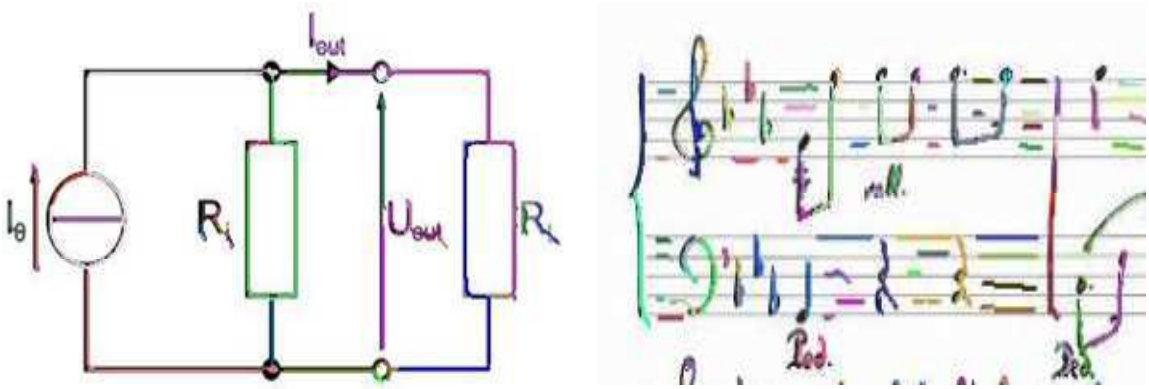
1 Conclusion générale

Nous avons présenté une contribution majeure du projet Graphem portant sur la mise en place d'une approche mixte (locale et globale) de caractérisation et de découpage des écritures du Moyen Age.

Contrairement aux approches purement globales ou locales, notre proposition présente plusieurs avantages. D'une part à partir d'une approche locale d'analyse des graphèmes et de leurs redondances, elle permet d'utiliser pleinement les informations morphologiques et les terminaisons des lettres que seule la recherche de l'axe médian que nous proposons permet de mettre en lumière. D'autre part une approche mixte produite à partir d'une décomposition en graphèmes issus d'une analyse locale de la dynamique de l'écriture et rassemblés en une signature globale caractéristique de toute une page manuscrite semble être une solution très efficace et discriminante de l'analyse des différents styles d'écriture paléographiques. La signature globale d'une page d'écriture a été calculée à partir d'un dictionnaire de formes représentant en différentes catégories, toute la diversité morphologique des graphèmes. A ce niveau, nous avons montré comment la coloration de graphe, engagée dans la construction des dictionnaires de formes, offrait une meilleure représentation des classes sans avoir recours à la moindre connaissance a priori du nombre de ces classes ou même de leurs tailles.

L'approche de décomposition des écritures manuscrites en graphèmes peut se généraliser à un grand panel de documents textuels : nos travaux sur la base de manuscrits contemporains *IAM* et *ICDAR 2011* l'attestent. Plus généralement, le principe d'une décomposition en traits est très répendu dans les images de documents et schémas graphiques, les partitions musicales et plus généralement les images de traits. Des travaux portant sur la décomposition en fragments linéaires (à partir d'une extraction de squelettes des formes) sur des images naturelles(ou des images de peintures) permettent d'envisager des pistes intéressantes d'application de nos travaux. La figure C.1 présente un exemple de décomposition d'un schéma graphique et de partitions musicales en graphèmes, en utilisant le même principe de décomposition que nous avons présenté dans le chapitre 2.

.



(a) (b)

Figure C.1. Décomposition, (a) schéma graphique, (b) partitions musicales, les traits en couleurs présentent les graphèmes extraient à partir de ces images

Afin de donner à notre approche une plus grande adaptabilité à la grande variété de formes des graphèmes et plus de souplesse vis-à-vis de certaines ambiguïtés, nous avons joint à la construction des dictionnaires de formes, une phase d'extraction automatique des poids et des seuils génériques spécifiques à chaque style à partir des algorithmes génétiques. Cette démarche nous a permis d'atteindre de meilleures classifications des graphèmes et par conséquent d'offrir une construction des dictionnaires de formes plus efficaces.

Nous avons démontré comment tous ces éléments ont contribué à une sélection simple de caractéristiques. A partir des résultats nous avons montré que les poids et le seuil générique sont adéquats aux bases de styles visuellement proches, ou comportant un nombre réduit de classes de styles (quelues dizaines maximum). Un apprentissage séparé par classes de styles s'avère en revanche indispensable dans le cas de bases de styles très différents ou en très grand nombre. Cette contribution représente aussi un nouveau point de vue dédié aux applications *CBIR*, nous l'avons développé dans le cadre du projet *GRAPHEM* et la méthodologie est applicable beaucoup plus largement sur des bases plus récentes et de contenus différents, comme l'attestent nos expérimentations.

2 Perspectives

Nous hiérarchisons les perspectives à court, moyen et long termes. Pour les perspectives à court terme, des travaux ont déjà été entrepris, mais nous n'avons pas à ce jour de résultats concrets à partir desquels nous pourrions tirer nos premières conclusions. Pour les perspectives à long terme, nous avons entrepris un plan d'expérimentations mais des développements importants resent à faire.

2.1 Les perspectives à courts termes

2.1.1 Vérité terrain pour les dictionnaires de formes

Durant nos travaux l'une des difficultés que nous avons eu dès le départ concerne l'absence d'une vérité terrain fiable sur laquelle nous aurions pu baser l'analyse des résultats produits par nos algorithmes de construction de dictionnaires de formes. Puisque chaque manuscrit se décompose d'une façon différente en raison du grand nombre de graphèmes présents sur une page d'écriture, cela conduit inévitablement à un clustering en un nombre différent de graphèmes. Comme solution à cette variabilité liée à la grande variété de textes anaysés, nous avons envisagé d'explorer la possibilité d'utiliser les *A*lgorithme *G*énétiques *I*nteractifs (AGI) *(Kim et Sho, [2000]),* dans le but d'avoir des solutions qui sont acceptables à la fois par les paléographes et les informaticiens, en prenant en considération deux mesures : la mesure visuelle des paléographes et notre mesure automatique basée sur l'uniformité intra classe et la disparité inter classes.

Les Algorithmes Génétiques Interactifs sont une extension des Algorithmes Génétiques où une interaction existe entre la fonction à optimiser et l'utilisateur, ce dernier guidant la méthode vers les solutions ayant les caractéristiques qu'il préfère. Les étapes d'évaluation et de sélection automatiques des individus présentent des individus à l'utilisateur qui sélectionne un certain nombre d'individus .Cela définit les conditions dans lesquelles les AGI peuvent être utilisés :

- Le problème à résoudre est tel que les préférences de l'utilisateur doivent être prises en compte et sont difficiles à formaliser par une fonction mathématique.
- Il est possible de visualiser (plus généralement représenter) les individus pour que l'utilisateur puisse les évaluer rapidement et exprimer ses préférences en sélectionnant les meilleurs individus. Cela implique notamment de limiter la population à un petit nombre d'individus.

Nous avons déjà codé un prototype à partir duquel les paléographes peuvent évaluer visuellement une classification de graphèmes. Quant à nous, cette approche nous permet de savoir quel sous ensemble de caractéristiques a été utilisé pour une classification qui a été choisie par les paléographes et si le choix paraît pertinent pour une optimisation de l'uniformité intra classe et la disparité inter classes. Un rapprochement entre les deux points de vue, celui des sciences humaines et celui des informaticiens est un enjeu important des applications pluridisciplinaires auxquelles nous contribuons.

2.1.2 Dictionnaire de formes universel

Durant nos travaux nous avons construit des dictionnaires de formes spécifiques à chaque scripteur indépendamment de la classe à laquelle il appartient. Nous envisageons d'élargir notre étude à l'utilisation des dictionnaires de formes universels qui ont donné de meilleurs taux de reconnaissance par rapport aux dictionnaires de formes spécifiques aux scripteurs *(Siddiqi, [2009]).*

2.1.3 Identification de scripteurs et reconnaissance de styles

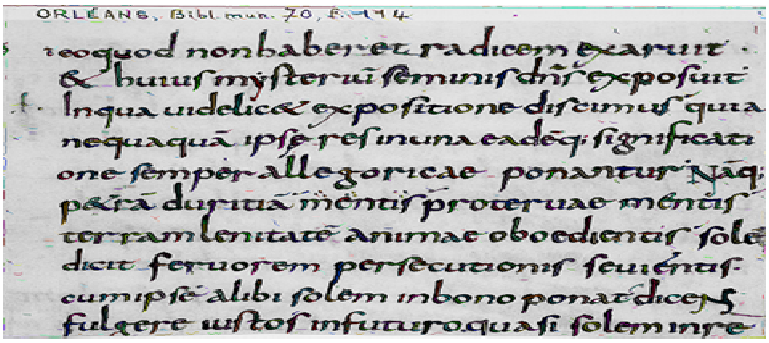
Dans le dernier chapitre de la thèse nous avons utilisé les dictionnaires de formes comme signature des écritures afin d'identifier les scripteurs (*ICDAR*, *IAM*) et classifier les manuscrits par style (*IRHT*, *Oxford*). Dans ces expériences nous avons utilisé trois distances pour calculer la similarité entre les manuscrits : distance de Hausdorff, distance de Hausdorff modifiée et la distance quadratique. Nous envisageons également de tester d'autres distances et plus précisément la distance Earth Mover's Distance qui est une métrique très intéressante dans le domaine des applications CBIR et qui a montré de bons niveaux de performance *(Rubner et al. [2000]).* Enfin des approches bayésiennes probabilistes nous semblent également des pistes prometteuses à suivre portant sur l'étude des probabilités d'apparitions de certaines formes redondantes, comme cela a déjà été initié par Schomaket et al dans *(Schomaker et al. [2007]).*

2.2 Les perspectives à moyen et long termes

2.2.1 Décomposition et extraction des caractéristiques

D'après les résultats d'identification de scripteur et de classification de manuscrits par styles, nous constatons que l'étape de décomposition qui utilise des règles spécifiques aux manuscrits médiévaux n'affecte pas les résultats d'identification ou de classification lors de leur application à des manuscrits contemporains. En effet notre méthode de décomposition a été appliquée à des manuscrits contemporains (*ICDAR 2011*) et nous avons pu enregistrer de bons résultats d'identification de scripteurs et de classification de manuscrits. L'influence du facteur « décomposition » par rapport au facteur « choix des caractéristiques » semble donc moindre.

Cette observation est également à associer au fait que les graphèmes sont classifiés indépendamment de leur ordre d'appartion dans la formation des mots.. Naturellement pour des applications très dépendantes de la structure interne des mots et ensemble de lettres (comme les applications de Word spotting ou de Word Retrieval) il devient indispensable de considérer l'ordre de formation des traits et d'appliquer des modifications à notre méthode la rendant robuste aux variations internes des mots (déformations même légères dans le processus d'écriture d'un mot) (figure C.2).



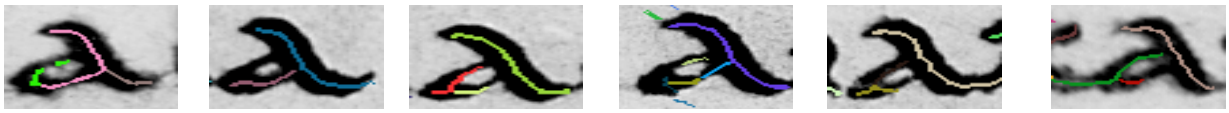


Figure C.2. Différentes décompositions de la lettre « a » dans un même manuscrit (IRHT)

Une modification de la méthode de décomposition devient nécessaire ici dans le but d'apporter une stabilité et une robustesse aux variations internes des écritures.

Nous avons commencé à explorer plusieurs pistes et plus précisément l'utilisation des tenseurs de vote qui ont prouvé leur robustesse au bruit, discontinuités, déformations qui se trouvent dans les documents *(Medioni et al. [2000])*. A partir des tenseurs nous pouvons extraire des caractéristiques comme la courbure et l'orientation et l'extraction du squelette peut se faire par une simple opération d'amincissement morphologique *(Loss et al. [2009])*. En restant dans le même esprit de l'étude de la dynamique de l'écriture, nous ouvrons une piste vers des applications nécessitant une très grande rigueur structurelle, comme les applications de Word spotting et Word retrieval.

Publications

Journaux nationaux

- Eglin.V, Gaceb.D, Daher.H, Bres.S, Vincent.N, *Outils d'analyse de la dynamique des écritures médiévales pour l'aide à l'expertise paléographique*, vol. 14, pp. 81-104, RDN, *2011*

Conférences nationales

- Daher.H, Eglin.V, Bres.S, Vincent.N, *Nouvelle approche d'extraction de l'axe médian dans les traits manuscrits : application à la constitution du code book des écritures*, CIFED, *2010.*
- Daher.H, Gaceb.D, Eglin.V, Bres.S, Vincent.N, *Décomposition des manuscrits anciens en graphèmes et construction des codes book basée sur la coloration de graphe*, CORESA, 2010.
- Daher.H, Gaceb.D, Eglin.V, Bres.S, Vincent.N, *Dictionnaire de formes par pondération de caractéristiques : application à l'analyse de manuscrits*, Dans Colloque International Francophone sur l'Écrit et le Document 2012 (CIFED), Bordeaux. pp. 1-16. mars *2012*
- Daher.H, Gaceb.D, Eglin.V, Vincent.N, S.Bres, *D'une pondération automatique des caractéristiques des graphèmes à la création des CodeBooks, un nouveau point de vue dédié aux applications CBIR*. RFIA, *2012*

Conférences internationales

- Daher.H, Eglin.V, Bres.S, Vincent.N, *A new approach for centerline extraction in handwritten strokes: an application to the constitution of a code book*, DAS, *2010.*
- Daher.H, Gaceb.D, Eglin.V, Bres.S, Vincent.N, *Ancient handwritings decomposition into graphemes and codebook generation based on Graph coloring*, IWFHR, 2010.
- Cloppet.F, Daher.H, Eglin.V, Emptoz.H, Exbrayat.M, Joutel.G, Lebourgeois.F, Martin.F, Moalla.I, Siddiqi.I, Vincent.N, *New Tools for Exploring, Analysing and Categorising Medieval Scripts*, Digital Medievalist, 2011.
- Daher.H, Gaceb.D, Eglin.V, Vincent.N, S.Bres, *Genetic Algorithm for Features Weighting and Automatic Parametrizing of the Classification Algorithm for Graphemes*, pp. 18-21, IPCV, *2011*.
- Daher.H, Gaceb.D, Eglin.V, Bres.S, Vincent.N, *Unsupervised categorization method of graphemes on handwritten manuscripts: application to style recognition*. Dans 19th Document Recognition and Retrieval Conference, pp. 0-8, DRR, *201*

Bibliographie

(Ablavsky et Stevens, [2003]) Ablavsky.V, Stevens.R, *Automatic Feature Selection with Applications to Script Identification of Degraded Documents*, pp. 750-754, ICDAR, 2003.

(Ahuja et Chuang, [1997]) Ahuja.N, Chuang.J, Shape *Representation using a Generalized Potential Field Model*, vol. 19, pp. 169-176, PAMI, 1997.

(Al-amri et al. [2003]) Al-amri.S, Kalyankar.N, Khamitkar.S, *Linear and Non-linear Contrast Enhancement Image*, vol. 10, IJCSNS, 2010.

(AlKhateeb, [2012]) AlKhateeb.J, *Offline Handwritten Arabic Digit Recognition Using Dynamic Bayesian Network*, ICCIT, 2012.

(Amaldi et Kann, [1998]) Amaldi.E, Kann.V, *On the approximation of minimizing non zero variables or unsatisfied relations in linear systems*, vol. 209, pp. 237-260, TCS, 1998.

(Atanasiu et al. [2011]) Atanasiu.V, Likforman-Sulem.L, Vincent.N, *Writer Retrieval - Exploration of a Novel Biometric Scenario Using Perceptual Features Derived from Script Orientation*, ICDAR, 2011.

(Attali, [1995]) Attali.D, *Squelettes et graphes de Voronoi 2D et 3D*, thèse, 1995.

(Aylward et Bullitt, [2002]) Aylward.S, Bullitt.E, *Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction*, vol. 2, pp. 61-75, MI, 2002.

(Babaguchi, [1990]) Babaguchi.N, *Connectionist model binarisation*, vol. 2, pp. 51-56, ICPR, 1990.

(Baird, [1987]) Baird.H, *The Skew Angle of Printed Documents*, pp. 14-21, CSPSE, 1987.

(Ball et Srihari, [2009]) Ball.G, Srihari.N, *Semi-Supervised Learning for Handwriting Recognition*, pp. 26-30, ICDAR, 2009.

(Banfield et Raftery, [1993]) Banfield.J, Raftery.A., *Model-based Gaussian and non-Gaussian clustering*, vol. 49, pp. 803-821, Biometrices, 1993.

(Beecks et al. [2009]) Beecks.C, Uysal.M, Seidl.T, *Signature Quadratic Form Distances for Content-Based Similarity*, pp.697-700, ICM, 2009.

(Beecks et al. [2010]) Beecks.C, Uysal.M, Seidl.T, *A Comparative Study of Similarity Measures for Content-Based Multimedia Retrieval*, pp. 1552-1557, ICME 2010.

(Belaid et Belaid, [1992]) Belaid.A, Belaid.Y, *Reconnaissance de formes : méthodes et applications*, 1992.

(Bensefia et al. [2002]) Bensefia.A, Nosary.A, Paquet.T, Heutte.L, *Writer identification by writer's invariants*, pp. 274-279, IWFHR, 2002.

(Bensefia et al. [2003]) Bensefia.A, Paquet.T, Heutte.L, *Information retrieval based writer identification*, pp. 946-950, ICDAR, 2003.

(Bensefia et al. [2005a]) Bensefia.A, Paquet.T, Heutte.L, *Handwritten document analysis for automatic writer recognition*, pp. 72-86, ELCVIA, 2005 (a).

(Bensefia et al. [2005b]) Bensefia, A., Paquet, T., Heutte, L, *A writer identification and verification system*, vol. 26, pp. 2080-2092, PRL, 2005 (b).

(Benzécri et Benzécri , [1980]) Benzécri.J, Benzécri.F, *Pratique de l'Analyse des données*, 1980.

(Bergevin et Bubel [2004]) Bergevin.R, Bubel.A, *Detection and characterization of junctions in a 2D image*, pp. 288-309, CVIU 2004.

(Bemsen, [1986]) Bernsen.J, *Dynamic Thresholding of Grey-Level Images*, ICPR, 1986.

(Bhardwaj et al. [2009]) Bhardwaj.A, Malgireddy.M, Setlur.S, Govindaraju.V, Sitaram.R, *Writer Identification in Offline Handwriting using Topic Models*, NIPS, 2009.

(Blum, [1964]) Blum.H, *A transformation for extracting new descriptors of shape*, pp. 362-380, MPSV 1964.

(Bordat, [1986]) Bordat.J, *Calcul pratique du treillis de Galois dune correspondance*, vol. 24, pp. 31-47, MISH, 1986.

(Boulehmi et al. [2008]) Boulehmi.H, Seddik.B, Kricha.A, *Ben Amara.N, Prétraitement de documents anciens*, CIFED, 2008.

(Boulétreau et al. [1995]) Boulétreau.V, vincent.N, Emptoz.H, *A writing qualification invariant towards line thickness and resolution changings*, pp. 325-329, ACCV, 1995.

(Boulétreau, [1997]) Boulétreau.V, *Vers un classement de l'écrit par des méthodes fractales*, thèse, 1997.

(Boulétreau et al. [1998]) Boulétreau.V, vincent.N, Sabourin.R, Emptoz.H, *Handwriting and signature: one or two personality identifiers?*, pp. 1758-1760, ICPR, 1998.

(Bribiesca, [1999]) Bribiesca.E, A new chain code, vol. 32, pp. 235–251, PR, 1999.

(Brink et al. [2012]) Brink.A, Smit.J, Bulacu.L, Schomaker.L, *Writer identification using directional ink-trace width measurements*, vol. 45, pp. 162-171, PR, 2012.

(Bui et al. [2011]) Bui.Q, Visani.M, Prum.S, Ogier.J, *Writer Identification Using TF-IDF for Cursive Handwritten Word Recognition*, pp.844-848, ICDAR, 2011.

(Bulacu et Schomaker, [2005]) Bulacu.M, Schomaker.L, A comparison *of clustering methods for writer identification and verification*, pp.1275-1279, ICDAR, 2005.

(Bulacu et Schomaker, [2006]) Bulacu.M, Schomaker.L, *Combining multiple features for text-independent writer identification and verification*, pp. 281-286, IWFHR, 2006.

(Bulacu et Schomaker, [2007]) Bulacu.M, Schomaker.L, *Text-independent writer identification and verification using textural and allographic features*, vol.29, pp.701-717,PAMI, 2007.

(Bullit et al. [2000]) Bullitt.E, Aylward.S, Bernard.E, Gerig.G, Womack.B, *Computer-assisted visualization of arteriovenous malformations on the home PC*, vol. 48, pp. 576-583, Neurosuergery, 2000.

(Busch et al. [2005]) Busch.A, Boles.W, Sridharan.S, *Texture for Script Identification*, vol. 27, pp. 1720-1732, PAMI, 2005.

(Camastra et Vinciarelli, [2001]) Camastra.F, Vinciarelli.A, *Cursive character recognition by learning vector quantization*, vol. 22, pp. 625–629, PRL, 2001.

(Can, [1993]) Can.F, *Incremental clustering for dynamic information processing*. vol. 11, pp. 143-164, IS, 1993.

(Cannon et al. [1999]) Cannon.M, Hockberg.J, Kelly.P, *Quality Assessment and Restoration of Typewritten Document Images*, vol.2, pp. 80-89, IJDAR, 1999.

(Canny, [1986]) Canny.J, *A computational approach to edge detection*, vol. 8, pp. 679-697, PAMI, 1986.

(Chan et al. [2000]) Chan.T, Sandberg.B, Vese.L, *Active contours without edges for vector-valued images*, vol. 11, pp. 130-141, JVCIR, 2000.

(Chanda et al. [2004]) Chanda.S, Sinha.S, pal.U, *Word-wise English Devnagari and Oriya Script Identification*, pp. 244-248, SPLASH 2004.

(Chang, [1995]) Chang.M, *Improved binarization algorithm for document image by histogram and edge detection*, vol. 2, pp. 636-639, ICDAR, 1995.

(Chapelle et Vapnik, [1999]) Chapelle.O, Vapnik.V, *Model Selection for Support Vector Machines*, pp. 230-236, NIPS, 1999.

(Chapelle et Zien, [2005]) Chapelle.O, Zien.A, *Semi-supervised classification by low density separation*, AIS 2005.

(Chapelle, [2005]) Chappelle.O, *Active learning for parzen windows classifier*, pp. 49–56, AIS, 2005.

(Chaudhury et heth, [1999]) Chaudhury.S, heth.R, *Trainable script identification strategies for Indian languages*, pp. 657–660, ICDAR, 1999.

(Cheeseman et Stutz, [1996]) Cheeseman.P, Stutz.J, *Bayesian Classification (AutoClass): Theory and Results*, pp. 153-180, KDD, 1996.

(Chenfg et al. [1998]) Cheng.H, Chen.J, Li.J, *Threshold selection based on fuzzy c-partition entropy approach*, vol. 31, pp. 857-870, PR, 1998.

(Cheng et al. [2002]) Cheng.J, Greiner.R, Kelly.J, Bell.D, Liu.W, *Learning Bayesian networks from data: An information-theory based approach*, vol.137, pp. 43–90, AI, 2002.

(Cheriet et Doré, [2006]) Cheriet.M, Doré.V, *Amincissement-sans-segmentation et re-haussement des images de niveau de gris par un filtre de chocs utilisant des champs de diffusion*, vol. 23, TS, 2006.

(Chigusa, [1992]) Chigusa.Y, *An image binarization system for composite pictures Circuits and Systems*, vol. 5, pp. 2292-2295, ISCAS, 1992.

(Cho et Kim, [2003]) Cho.S, Kim.J, *Bayesian network modeling of hangul characters for on-line handwriting recognition*, ICDAR, 2003.

(Choi et al. [2003]) Choi.W, Lam.K, Siu.W, *Extraction of the Euclidean skeleton based on a connectivity criterion*, vol. 36, PR, 2003.

(Chouaib et al. [2009]) Chouaib.H, vincent.N, Cloppet.F, Tabbone.S, *Generic Feature Selection and Document Processing*, pp. 356-360, ICDAR, 2009.

(Chung et Sapiro, [2000]) Chung.D, Sapiro.G, *Segmentation-Free Skeletonization of Gray-Scale Images via PDE*, Vol. 2, pp 927-930, ICIP, 2000.

(Cocquerez et Philipp, [1995]) Cocquerez.J, Philipp.S, *Analyse d'images : Filtrage et Segmentation*, 1995.

(Cohn et al. [2003]) Cohn.D, Caruana.R, McCallum.A, *Semi-supervised clusteringwith user feedback*, Technical Report, 2003.

(Cover et Hart, [1967]) Cover.T, Hart.P, *Nearest neighbor pattern classification*, vol. 13, pp. 21–27, IT, 1967

(Cover et Van Campenhout, [1997]) Cover.T, Van Campenhout.J, *On the possible orderings in the measurement selection problem*, vol.7, pp. 657–661,SMC, 1997.

(Cunnigha et Denaly, [2007]) Cunningham.P, Delany.S, *k-Nearest Neighbour Classifiers*, Technical Report, 2007.

(Danielsson, [1980]) Danielsson.P, *Euclidean Distance Mapping*, vol. 14, pp. 227-248, CGIP, 1980.

(Dargenton et al. [1991]) Dargenton.P, vincent.N, Emptoz.H, *Segmentation de l'écriture cursive à l'aide de la transformée de Fourier*, Vol. 2, pp. 707-712, RFIA, 1991.

(Das et al. [2011]) Das.M, Rani.D, Reddy.C, Govardhan.A, *Script identification from Multilingual Telugu, Hindi and English Text Documents*, vol. 1, IJWBC, 2011.

(Dash et Liu, [1997]) Dash.M, Liu.H, *Feature selection for classifications*, pp. 131-156, IDA 1997.

(David, [1991]) Davis.T, *The Handbook of Genetic Algorithms*, 1991.

(Dawson et Wilby, [2001]) Dawson.C, Wilby.R, *Hydrological modelling using artificial neural networks*, vol. 25, pp. 80-108, PPG, 2001.

(Dempster et al. [1977]) Dempster.A, Laird.N, Rubin.D, *Maximum likelihood from in complete data using the EM algorithm*, vol. 39, JRSS 1977.

(Dhanya et al. [2002]) Dhanya.D, Ramkrishnan.A, Pati.P, *Script Identification in Printed Bilingual Documents*, vol. 27, pp. 73-82, Sadhana, 2002.

(Diday, [1971]) Diday.E, *La méthode des nuées dynamiques*, vol. 19, pp. 19-34, SA, 1971.

(Diday, [1972]) Diday.E, Simon.J, *Clustering analysis*. pp. 47-94, DPR, 1976.

(Drira et al. [2006]) Drira.F, Lebourgeois.F, Emptoz.H, *Séparation recto/verso des images de documents anciens par une approche «aveugle»*, pp. 199-204, CIFED, 2006.

(Eglin et Volpilhac-Auger, [2004]) Eglin.V, Volpilhac-Auger.C, *Caractérisation multiéchelle des tracés manuscrits en vue de la catégorisation de scripteurs*, CIFED, 2004.

(Eglin et al. [2006]) Eglin.V, Lebourgeois.F, Bres.S, Emptoz.H, Y. Leydier, Moalla. I, F. Drira, *Computer assistance for Digital Libraries: Contributions to Middle-ages and Authors' Manuscripts exploitation and enrichment*, pp.265-280, DIAL, 2006.

(Eklund et Hoang, [2002]) Eklund.P, Hoang.A, *A Performance Survey of Public Domain Machine Learning Algorithms*, Technical Report, 2002.

(Elgammal et Ismail, [2001]) Elgammal.A, Ismail.M, *Techniques for Language Identification for Hybrid Arabic-English Document Images*, pp. 1100-1104, ICDAR, 2001.

(Emmanuel, [1974]) Emmanuel.P, *Paléographie et méthodologie. Vers l'analyse scientifiques des écritures médiévales*, pp. 101-110, 1974.

(Esquef, [2002]) Esquef.I, M. Albuquerque, *Nonextensive entropic image thresholding*, pp. 402, SIBGRA, 2002.

(Ester, [1996]) Ester.M, Kriegel.H, Sander.S, Xu.X, *A density-based algorithm for discovering clusters in large spatial databases with noise*, pp. 226-231, KDD, 1996.

(Fan et al. [1998]) Fan.K, Chen.D, Wen.M, *Skeletonization of Binary Images with Nonuniform Width via Block Decomposition and Contour Vector Matching*, vol. 31, pp. 823–838, PR, 1998.

(Fisher, [1987]) Fisher.D, *Knowledge acquisition via in- cremental conceptual clustering*, vol. 2, pp. 139-172, ML, 1987.

(Fisher, [1958]) Fisher.W, *On grouping for maximum homogeneity*, vol. 53, pp. 789-798, JASA, 1958.

(Fisher, [1995]) Fisher.Y, *Fractal image compression*, Springer, 1995.

(Florack et al. [1992]) Florack.L, Romeny.B, Koenderink.J, Viergever.M, *Scale and the differential structure of images*, vol. 10, pp. 376-388, IVC 1992.

(Fomes et al. [2008]) Fornes.A, Llados.J, Sanchez.G, Bunke.H, *Writer Identification in Old Handwritten Music Scores*, DAS, 2008.

(Fort et al. [2002]) Fort.J, Letremy.P, Cottrell.M, *Advantages and drawbacks of the Batch Kohonen algorithm*, pp. 223-230, ESANN, 2002.

(Frale et Raftery, [1998]) Fraley.C, Raftery.A, *How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis*, Technical Report, 1998.

(Frangi et al. [1998]) Frangi.A, Niessen.W, Vincken.K, Viergever.M, *Multiscale vessel enhancement filtering*, vol. 1496, pp. 130–137, MICCAI, 1998.

(Funahashi, [1998]) Funahashi.K, *Multilayer neural networks and Bayes decision theory*, vol. 11, pp. 209–213, NN, 1998.

(Gabor, [1946]) Gabor.D, *Theory of communication. Journal of the Institute of Electrical Engineers*, pp. 429–457, 1946.

(Gaceb et al. [2007]) Gaceb.D, Eglin.V, Lebourgeois.F, Emptoz.H, *A New Pyramidal Approach for the Address Block Location Based on Hierarchical Graph Coloring*, pp. 1276-1288, ICIAR, 2007.

(Gaceb et al. [2008]) Gaceb.D, Eglin.V, Emptoz.H, *Improvement of postal mail sorting system*, IJDAR, 2008.

(Gaceb et al. [2009]) Gaceb.D, Eglin.V, Lebourgeois.F, Emptoz.H, *Graph b-Coloring for Automatic Recognition of Documents*, pp. 261-265, ICDAR, 2009.

(Gatos et al. [2006]) Gatos.B, Pratikakis.I, Perantonis.S, *Adaptive degraded document image binarization*, vol. 39, pp. 317–327, PR, 2006.

(Ghiasi et Safabakhsh, [2010]) Ghiasi.G, Safabakhsh.R, *An Efficient Method for O ine Text Independent Writer Identication*, pp.1245-1248, ICPR, 2010.

(Ghosh et Dude, [2009]) Ghosh.D, Dube.T, Shivaprasad.A, *Script Recognition a review*, PAMI, 2009.

(Gilliam et al. [2010]) Gilliam.T, Wilson.R, Clark.J, *Scribe Identification in Medieval English Manuscripts*, pp. 1880-1883, ICPR, 2010.

(Glasbey, [1993]) Glasbey.C, *An analysis of histogram-based thresholding algorithms*, vol. 55, pp. 532-537, CVGIP, 1993.

(Goldberg, [1989]) Goldberg.D, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

(Gonzalez et Woods, [2002]) Gonzalez.A, Woods.R, *Digital Image Processing*, Addison-Wesley, 2002.

(Gowda et Diday, [1992]) Gowda.K, Diday.E, *Symbolic clustering using a new dissimilarity measure*, vol. 22, pp. 368–378, SMC, 1992.

(Griogorishin et al. [1998]) Grigorishin.T, G. Abdel-Hamidet, Yang.H, *Skeletonisation: An Electrostatic Field Based Approach*, vol. 1, pp. 163-177, PAA, 1998.

(Guo et Zhao, [2010]) Guo.H, Zhao.J, Chinese *Minority Script Recognition Using Radial Basis Function Network*, vol. 5, pp. 927-934, JCP, 2010.

(Hagedoom et al. [2000]) Hagedoorn.M, Overmars.M, Veltkamp.R, *A new visibility partition for affine pattern matching. In Discrete Geometry for Computer Imagery*, pp. 358-370, DGCI, 2000.

(Hamza et al. [2005]) Hamza.H, Smigiel.E, Belaid.E, *Neural based binarization techniques*, vol. 1, pp. 317-32, ICDAR, 2005.

(Hangarge et Dhandra, [2009]) Hangarge.M, Dhandra.B, *Script Identification in Indian Document Images based on Directional Morphological Filters*, vol. 2, IJRTE, 2009.

(Haralick et al. [1973]) Haralick.R, Shanmugam.K, Dinstein.I, Textural Features for Image Classification, vol. 3, SMC, 1973.

(Haralick et Shapiro, [1985]) Haralick.M, L.Shapiro.L, Image segmentation techniques, vol. 29, pp. 100-132, CVGIP, 1985.

(Hartigan, [1975]) Hartigan.J, *Clustering algorithms*, JohnWiley and Sons., 1975.

(He et al. [2005]) He.J, Do.Q, Downton.A, Kim.J, *A comparison of binarization methods for historical archive documents*, vol. 1, pp. 538-542, ICDAR, 2005.

(Hinds et al. [1990]) Hinds.S, Fisher.J, D'Amato.D, *A Document Skew detection method using run-length encoding and the Hough transform*, vol. 1, pp. 464-468, ICPR, 1990.

(Hochberg et al. [1997]) Hochberg.J, Kelly.P, Thomas.T, Kerns.L, *Automatic Script Identification from Document Images Using Cluster-based Templates*, vol. 19, pp. 176-181, PAMI, 1997.

(Huang et al. [2008]) Huang.C, Yang.D, Chuang.Y, *Application of wrapper approach and composite classifier to the stock trend prediction*, vol. 34, pp. 2870-2878, ESA, 2008.

(Huang et Wang, [2006]) Huang.L, Wang.J, A GA-based *feature selection and parameters optimization for support vector machines*, vol. 31, pp. 231-240, ESA 2006.

(Hubert, [1973]) Hubert.L, *Monotone invariant clustering procedures*, vol. 30, pp. 47-63, Psychometrika, 1973.

(Howlett et Jain, [2001]) Howlett.R, Jain.L, *Radial basis function networks 2 : new advances in design*, vol. 67, SFSC, 2001.

(Imdad et al. [2007]) Imdad.A, Bres.S, Eglin.V, *Writer identification using steered hermite features and svm*, vol. 2, pp. 839–843, ICDAR, 2007.

(Isaacman et al. [2011]) Isaacman.S, Becker.R, Caceres.R, Kobourov.S, Martonosi.M, Rowland.J, Varshavsky.A, *Identifying important places in people's lives from cellular network data*, ICPC, 2011.

(Jaeger et al. [2005]) Jaeger.S, Ma.H, Doermann.D, *Identifying Script on Wordlevel with Informational Confidence*, vol. 1, pp. 416-420, ICDAR, 2005.

(Jain et al. [1996]) Jain.A, Zhong.Y, *Page Segmentation Using Texture Analysis, Pattern Recognition*, vol. 29, pp. 743-770, May 1996.

(Jain et al. [1999]) Jain.A, Murty.M, Flynn.P, *Data clustering: a review*, vol. 31, pp. 264–323, CS, 1999.

(Jain et Duin, [2000]) Jain.R, Duin.J, *Statistical Pattern Recognition: A Review*, vol. 22, pp. 4-37, PAMI, 2000.

(Jain et Doermann, [2011]) jain.R, Doermann.D, *Offline Writer Identification using K-Adjacent Segments*, ICDAR, 2011.

(Jalba et al. [2006]) Jalba.A, Wilkinson.H, Roerdink.J, *Shape representation an recognition through morphological curvature scale spaces*, vol. 15, pp. 331-341, IP, 2006.

(Jambu et Lebeaux, [1978]) Jambu.M, Lebeaux.M, *Classification automatique pour l'Analyse des données*, 1978.

(Joshi et al. [2006]) Joshi.G, Garg.S, Sivaswamy.J, *Script Identification from Indian Documents*, pp. 255-267, LNCS, 2006.

(Joutel et al. [2008]), Joutel.G, Eglin.V, Emptoz.H, *A complete pyramidal geometrical scheme for text based image description and retrieval*, pp. 471-480, ICISP, 2008.

(Joutel, [2009]), Joutel.G, *Analyse multirésolution des images de documents manuscrits : Application à l'analyse de l'écriture*, thèse, 2009.

(Kng et Kim, [2004]) Kang.K, Kim.J, *Utilization of Hierarchical, Stochastic Relationship Modeling for Hangul Character Recognition*, Vol. 26, pp. 1185-1196, PAMI, 2004.

(Kapur et al. [1985]) Kapur.J, Sahoo.P, Wong.A, *A new method for gray-level picture thresholding using the entropy of the histogram*, vol. 29, pp. 273-285, GMIP, 1985.

(Kass et al. [1987]) Kass.M, Witkin.A, Terzopoulos.D, *Snakes: Active contour models*, pp. 259–268, ICCV, 1987.

(Kato et Yasuhara, [2000]) Kato.Y, Yasuhara.M, *Recovery of drawing order from single-stroke handwriting images*, vol. 22, pp. 938-949, PAMI, 2000.

(Kau et al. [2011]) Kaur.M, Kaur.J, Kaur.J, *Survey of Contrast Enhancement Techniques based on Histogram Equalization*, vol. 2, IJACSA, 2011.

(Kavallieratou et Stathis, [2006]) Kavallieratou.E, Stathis.S, *Adaptive Binarization of Historical Document Images*, vol. 3, pp. 742-745, ICPR, 2006.

(Khashman et Sekeroglu, [2008]) Khashman.A, B. Sekeroglu.B, *Document Image Binarisation Using a Supervised Neural Network*, vol. 18, pp. 405-418, IJNS, 2008.

(Khurshid et al. [2009]) Khurshid.K, Siddiqi.I, Faure.C, vincent.N, *Comparison of Niblack inspired Binarization methods for ancient documents*, ICDAR, 2009.

(Kim et Cho, [2000]) Kim.H, Cho.S, *Application of interactive genetic algorithm tofashion design*, vol. 13, pp. 635–44, EAAI, 2000.

(Kim et al. [2001]) Kim.J, Kim.L, Hwang.S, *An advanced contrast enhancement using partially overlapped sub-block histogram equalization*, vol. 11, pp. 475-484, CSVT, 2001.

(Kim et Lee, [1998]) Kim.S, Lee.S, *Gray-scale nonlinear shape normalization method for handwritten oriental character recognition*, vol. 12, pp. 81-95, IJDAR, 1998.

(Kimmel et al. [1995]) Kimmel.R, Shaked.D, Kiryati.N, *B*ruckstein.A, *Skeletonization via Distance Maps and Level Sets*, vol. 62, pp. 382-391, CVIU, 1995.

(Koenderink, [1984]) Koenderink.J, *The structure of images*, vol. 50, pp. 363-370, BC, 1984.

(Kohavi et al. [1995]) Kohavi.R, Langley.P, Yun.Y, *Heuristic search for feature weights in instance-based learning*, 1995.

(Kohavi et Quinlan, [2002]) Kohavi.R, Quinlan.J, *Decision-tree discovery*, pp. 267-276. Handbook of Data Mining and Knowledge Discovery, 2002.

(Köhler, [1981]) Köhler.R, *A segmentation system based on thresholding*, vol. 15, pp. 319-338, GMIP, 1981.

(Kohonen, [1982]) Kohonen.T, *Self-organized formation of topologically correct feature maps*, vol. 3, pp. 59-69, BC, 1982.

(Kohonen, [1988]) Kohonen.T, *Self-Organization and Associative Memory*, Springer, 1988.

(Kourtis et Stamatos, [2011]) Kourtis.I, Stamatatos.E, *Author Identification Using Semi-supervised Learning-Notebook for PAN*, CLEF, 2011.

(Kukuck, [1980]) kuckuck.W, *Writer recognition by spectra analysis*, pp. 1-3, ICSE, 1980.

(Kumar et al. [2003]) Kumar.R, Chaitanya.V, Jawahar.C, *A Novel Approach to Script Separation*, pp. 289-292, ICAPR, 2003.

(Lakshmi et Punithavalli, [2010]) Lakshmi.J, Punithavalli.M, *2D Shape Reconstruction Based on Combined SkeletonBoundary Features*, vol. 4, IJIP, 2010.

(Land et Doig, [1960]) Land.A, Doig.A, *An automatic method of solving discrete programming problems*, vol. 28, pp. 497–520, Econometrica, 1960.

(Lebourgeois et Hemptoz, [2007]) Lebourgeois.F, Emptoz.H , *Skeletonization by gradient regularization and diffusion*, pp. 1118-1122, ICDAR, 2007.

(Lee et al. [1978]) Lee.C, Slagle.J, Mong.C, *Towards automatic auditing of records*, vol. 4, pp. 441-448, SE 1978.

(Lee et al. [1996a]) Lee.D, Nohl.C, Baird.H, *Language Identification in Complex, Unoriented, and Degraded Document Images*, pp. 76-98, IAPR, 1996 (a).

(Lee et Kim, [1995]) Lee.S, Kim.J, *Multi-lingual, Multi-font, Multi-size Large-set Character Recognition Using Self-organizing Neural Network*, vol. 1, pp. 28-33, ICDAR, 1995.

(Lee et al. [1996b]) Lee.W, Lee.D, Park.H, *A New Methodology for GrayScale Character Segmentation and Recognition*, vol. 18, pp. 1045-1050, PAMI, 1996 (b).

(Legault et Suen, [1992]) Legault.R, Suen.C, *A comparison of methods of extracting curvature features*, ICPR, 1992.

(Levina et Bickel, [2001]) Levina.E, Bickel.P, *The EarthMover's Distance is the Mallows Distance: Some Insights from Statistics*, pp. 251-256, ICCV, 2001.

(Levine et Nazif, [1985]) Levine.M, Nazif.A, *Dynamic measurement of computer generated image segmentations*, pp. 155-164, PAMI, 1985.

(Li et Lee, [1993]) Li.C, Lee.C, *Minimum Cross Entropy Thresholding*, vol. 26, pp. 616-626, PR, 1993.

(Li et Guo, [2008]) Li.Y, Guo.L, *TCM-KNN scheme for network anomaly detection using feature-based optimizations*, pp. 2103-2109, SAC, 2008.

(Lin et al. [2011]) Lin.X, Guo.C, Chang.F, *Classifying Textual Components of Bilingual Documents with Decision-Tree Support Vector Machines*, ICDAR, 2011.

(Lins et al. [1994]) Lins.R, Guimaraes Neto.M, França Neto.L, Galdino Rosa.L, *An Environment for Processing Images of Historical Documents*, vol. 40, pp. 939-942, MM, 1994.

(Liu et al. [2008]) Liu.F, Peng.X, Wang.T, Lu.S, *A density-based approach for text extraction in images*, ICPR, 2008.

(Loss et al. [2009]) Loss.A, Bebis.G, Parvin.B, *Tunable Tensor Voting for Regularizing Punctate Patterns of Membrane-Bound Protein Signals*, ISBI, 2009.

(Ma et Doermann, [2003]) Ma.H, Doermann.D, *Gabor Filter Based Multi-class Classifier for Scanned Document Images*, pp. 968-972, ICDAR, 2003.

(Maaten et Postma, [2005]) Maaten.L, Postma.E, *Improving automatic writer identification*, pp. 260–266, CAI, 2005.

(Maio et Maltoni, [1997]) Maio.D, Maltoni.D, *Direct Gray-Scale Minutiae Detection in Fingerprints*, vol. 19, pp. 27-40, PAMI, 1997.

(Mali et Mitra, [2003]) Mali.K, Mitra.S, *Clustering and its validation in a symbolic framework*, vol. 24, pp. 2367-2376, PRL, 2003.

(Mandelbrot, [1975]) Mandelbrot.B, *Les objets fractals*, 1975.

(Marr et Hildreth, [1980]) Marr.D, Hildreth.E, *Theory of edge detection*, RS, 1980.

(Martens, [1990]) Martens.J, *The Hermite transform – Theory*, vol. 38, pp. 1595- 1606, ASS, 1990.

(Matula, [1972]) Matula.D, *K-Components, Clusters, and Slicings in Graphs*, vol. 22, pp. 459-480, SIAM, 1972.

(Mayer et al. [1998]) Mayer.H, Ivan.L, Baumgartner.A, *Multi-Scale and Snakes for Automatic Road Extraction*, pp. 720-733, CCV, 1998.

(Mazzei et al. [2011]) Mazzei.A, Kaplan.F, Dillenbourg.P, *Extraction and Classification of Handwritten Annotations*, 2011.

(McInerney et al. [1996]) McInerney.T, Terzopoulos.D, *Deformable models in medical image analysis: A survey*, vol. 1, pp. 91–108, MIA, 1996.

(Medioni et al. [2000]) Medioni.G, Lee.M, Tang.C, *A Computational Framework for Segmentation and Grouping*, Elsevier, 2000.

(Meyer et Maragos, [1999]) Meyer.F, Maragos.P, *Multiscale Morphological Segmentations Based on Watershed, Flooding, and Eikonal PDE*, pp. 351-362, SSTCV, 1999.

(Meyer, [1989]) Meyer.Y, Ondelettes, *filtres miroirs en quadrature et traitement numérique de l'image*, 1989.

(Mitchell, [1996]) Mitchell.M, *An introduction to genetics algorithms*, MIT, 1996.

(Miura et al. [1997]) Miura.K, Sato.R, Mori.S, *A method of extracting curvature features and its application to handwritten character recognition*, pp. 450–454, ICDAR, 1997.

(Moalla et al. [2002]) Moalla.I, Elbaati.A, Alimi.A, Benhamadou.A, *Extraction of Arabic Text from Multilingual Documents*, ICSMC, 2002.

(Moalla et al. [2004]) Moalla.I, Alimi.A, Benhamadou.A, *Extraction of Arabic Words from Multilingual Documents*, CAISC, 2004.

(Moalla et al. [2006]) Moalla.I, Alimi.A, Lebourgeois.F, Emptoz.H, *Image Analysis for Palaeography Inspection*, pp. 303-311, DIAL, 2006.

(Moalla et al. [2006]) Moalla.I, Lebourgeois.F, Emptoz.H, Alimi.A, *Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography*, pp. 25-37, DAS, 2006.

(Moalla et al. [2009]) Moalla.I, *Caratérisation des écritures médiévales par des méthodes statistiques basées sur les cooccurrences*, thèse, 2009.

(Moghaddam et Cheriet, [2009]) Moghaddam.R, Cheriet.M, *Rsldi: restoration of singlesided low-quality document images*, vol. 42, pp. 3355-3364, PR, 2009.

(Mokhtar et al. [2009]) Mokhtar.N, Harun.N, Mashor.M, Roseline.H, Mustafa.N, Adollah.R, *Image Enhancement Techniques Using Local, Global, Bright, Dark and Partial Contrast Stretching For Acute Leukemia Images*, WCE, 2009.

(Mokhtarian et al. [1996]) Mokhtarian.F, Abbasi.S, Kittler.J, *Efficient and robust retrieval by shape content through curvature scale space. In Image Databases and Multi-Media Search*, pp. 35-42, IDB-MMS, 1996.

(Nakache et Confais, [2005]) Nakache.J, Confais.J, *Approche pragmatique de la classifciation*, Editions Technip, 2005.

(NG et Han, [2002]) NG.R, Han.J, *CLARANS : A method for clustering objects for spatial data mining*, vol. 14, pp. 1003-1016, KDE, 2002.

(Nguyen et al. [2008]) Nguyen.T, Tabbone.S, Terrades.O, *Symbol descriptor based on shape context and vector model of information retrieval*, p. 191–197, DAS, 2008.

(**Niblack, [1986])** Niblack.W, *An Introduction to Digital Image Processing*, pp. 115-116, Prentice Hall, 1986.

(Nosary et al. [2004]) Nosary.A, Heutte.L, Paquet.T, *Unsupervised writer adaptation applied to handwritten text recognition*, vol. 37, pp. 385–388, PR, 2004.

(Osher et Sethian, [1988]) Osher.S, Sethian.J, *Fronts propagation with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations*, vol. 79, pp. 12–49, JCP, 1988.

(Otsu, [1979]) Otsu.N, *A threshold selection method from gray-level histograms*, vol. 9, pp. 62-66, SMC 1979.

(Pal et al. [1999]) pal.U, Chaudhuri.B, *Script Line Separation from Indian Multi-script Documents*, pp. 406-409, ICDAR, 1999.

(Pal et al. [2002]) pal.U, Chaudhuri.B, *Identification of Different Script Lines from Multi-script Documents*, vol. 20, no. 13-14, pp. 945-954, IVC, 2002.

(Pan et al. [2005]) Pan.W, Suen.C, Bui.T, *Script Identification Using Steerable Gabor Filters*, vol. 2, pp. 883-887, ICDAR, 2005.

(Pareti et Vincent, [2006]) Pareti.R, vincent.N, *Global method based on pattern occurrences for writer identification*, IWFHR, 2006.

(Pareti et Vincent, [2008]) Pareti.R, vincent.N, *Indexation de lettrine par une méthode hybride*, CIFED, 2008.

(Patil et al. [2002]) Patil.S, Subbareddy.N, *Neural Network Based System for Script Identification in Indian Documents*, vol. 27, pp. 83-97, Sadhana, 2002.

(Peake et Tan, [1998]) Peake.G, Tan.T, *Script and Language Identification from Document Images*, pp. 97-104, LNCS, 1998.

(Pei et al. [2004]) Pei.S, Zeng.Y, Chang.C, *Virtual restoration of ancient Chinese paintings using color contrast enhancement and lacuna texture synthesis*, vol. 13, pp. 416–429, 2004.

(Pervouchine et Leedham, [2005]) Pervouchine.V, Leedham.G, *Document examiner feature extraction: Thinned vs. skeletonised handwriting images*, pp. 1-6, TENCON, 2005.

(Pervouchine et Leedham, [2007]) Pervouchine.V, Leedham.G, *Study of structural features of handwritten grapheme 'th' for writer identification*, pp. 417-422, IAS, 2007.

(Pereti et al. [2003]) Peteri.R, Celle.J, Ranchin.T, *Detection and extraction of road networks from high resolution satellite images*, vol. 1, pp. 301-304, ICIP, 2003.

(Phyu, [2009]) Phyu.T, *Survey of Classification Techniques in Data Mining*, IMECS, 2009.

(Pizer et al. [1984]) Pizer.S, Zimmerman.J, Staab.E, *Adaptive grey level assignment in CT scan display*, vol. 8, pp. 300–305, JCAT, 1984.

(Pos et al. [2006]) Poz.A, ZaninetG.R, do Vale.M, *Automated extraction of road network from medium and high-resolution images*, vol. 16, pp. 239-248, PRIA, 2006.

(Pratikakis et al. [2011]) Pratikakis.I, Gatos.B, Ntirogianni.K, *Document Image Binarization Contest*, ICDAR, 2011.

(Pudil et al. [2005]) Pudil.P, Novovicova.J, Kittler.J, *Floating search methods infeature selection*, vol. 15, pp.1119–1125, PRL, 1995.

(Qui, [2004]) Qi.Y, *Fingerprint Ridge Line Reconstruction*, pp. 211-220, IIP, 2004.

(Quinlan, [1979]) Quinlan.J, *Discovering rules by induction from large collections of examples*, pp. 168–201, ESMEA, 1979.

(Quinlan, [1993]) Quinlan.J, *C4.5: Programs for Machine Learning*, *Morgan Kaufman Publishers*, 1993.

(Rafat et Soryani, [2006]) Rafat.N, Soryani.M, *Application of Genetic Algorithms to Feature Subset Selection in a Farsi OCR*, CISE 2006.

(Rajput et Anita, [2010]) Rajput.G, Anita.H, *Handwritten Script Recognition using DCT and Wavelet Features at Block Level*, vol. 3, pp. 158–163, IJCA, 2010,

(Ramer, [1972]) Ramer.U, *An iterative procedure for the polygonal approximation of plane curves*, vol. 1, pp. 244-256, CGIP, 1972.

(Raymer et Punch, [2000]) Raymer.M, Punch.W, Goodman.E, Kuhn.L, Jain.A, *Dimensionality reduction using genetic algorithms*, vol.4, pp. 164-171, E, 2000.

(Ronse et Devijver, [1984]) Ronse.C, Devijver.P, *Connected Components in Binary Images: The Detection Problem*, John Wiley & Sons, New York, 1984.

(Rosenfield et Pfalz, [1968]) Rosenfeld.A, Pfalz.J, *Distance Functions on Digital Pictures*, vol. 1, pp. 33-61, PR, 1968

(Rousseau et al. [2004]) Rousseau.L, Anquetil.E, Camillerapp.J, *Reconstitution du parcours du tracé manuscrit hors-ligne de caractères isolés*, CIFED, 2004.

(Roy et Pal, [2004]) Roy.K, pal.U, *Chaudhuri.B, Address Block Location and Pin Code Recognition for Indian Postal Automation*, pp. 5-9, WVGIP, 2004.

(Roy et al. [2005]) Roy.K, Vajda.S, pal.U, Chaudhuri.B, A. Belaid, *A System for Indian Postal Automation*, vol. 2, pp. 1060-1064, ICDAR, 2005.

(Rubner et al. [1997]) Rubner.Y, Guibas.L, Tomasi.C, *The earth mover's distance, multi-dimensional scaling*, and color-based image retrieval, pp. 661–668, ARPA, 1997.

(Russel et Norvig, [2003]) Russell.S, Norvig.P, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2003.

(Saeys et al. [2007]) Saeys.Y, Inza.I, Larranaga.P, *A review of feature selection techniques in bioinformatics*, vol. 23, pp. 2507-2517, Bioinformatics, 2007.

(Sahoo et al. [1988]) Sahoo.P, Soltani.S, Wong.A, Chen.Y, *A survey of thresholding techniques*, vol. 41, pp. 233-260, CGIP, 1988.

(Said et al. [2000]) Said.H, Tan.T, Baker.K, *Personal Identification Based on Handwritting*, vol. 33, pp 149-160, PR, 2000.

(Sait et Youssef, [1999]) Sait.S, Youssef.H, *General iterative algorithms for combinatorial optimization*, CS, 1999.

(Salman, [2006]) Salman.N, *Image Segmentation Based on Watershed and Edge Detection Techniques*, vol. 3, pp. 104-110, IAJIT, 2006.

(Sauvola, [1997]) Sauvola.J, *Adaptive document binarisation*, vol. 1, pp. 147-152, DAS, 1997.

(Sauvola et Pietikäinen, [2000]) Sauvola.J, Pietikäinen.M, *Adaptive document image binarization*, vol. 33, pp. 225-236, PR, 2000.

(Savakis, [1998]) Savakis.A, *Adaptive document image thresholding using foreground and background clustering Image Processing*, vol. 3, pp. 785-789, ICIPP, 1998.

(Scassellati et al. [1994]) Scassellati.B, Slexopoulos.S, Flickner.M, *Retrieving images by 2D shape: a comparison of computation methods with human perceptual judgments*, pp. 2–14, SPIE, 1994.

(Schaefer, [2002]) Schaefer.G, *Compressed domain image retrieval by comparing vector quantization codeb*ooks, Proc. SPIE, vol. 4671, p. 959 , 2002

(Schimdt, [1989]) Schmidt.M, *Some examples of algorithm analysis in computer geometry by means of mathematical morphology techniques*, pp. 225-246, GR, 1989.

(Schomaker et al. [2004]) Schomaker.L, Bulacu.M, *Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script*, vol. 26, pp. 787-798, PAMI, 2004.

(Schomaker et al. [2007]) Schomaker.L, Franke.K, Bulacu.M, *Using codebooks of fragmented connected-component contours in forensic and historic writer identification*, vol.28, pp.719-727, PRL, 2007.

(Schomaker, [2007]) Schomaker.L, *Writer identification and verification*, pp. 247-264, Springer, 2007.

(Schroeder, [1976]) Schroeder.A, *Analyse d'un mélange de distributions de probabilité de même type*, vol. 24, pp. 39-62, RSA, 1976.

(Séropian et al. [2003]) Séropian.A, Grimaldi.M, vincent.N, *Writer identification based on the fractal construction of a reference base*, pp. 1163-1167, ICDAR, 2003.

(Séropian et al. [2004]) Séropian.A, Grimaldi.M, vincent.N, Differentiation of alphabets in handwritten texts, ICPR, 2004.

(Shahabi et Rahmati, [2006]) Shahabi.F, Rahmati.M, *Comparison of Gabor- based features for writer identification of Farsi/Arabic handwriting*, IWFHR, 2006.

(Shanmugavadivu et Balasubramanian, [2010]) Shanmugavadivu.P, Balasubramanian.K, *Image Inversion and Bi Level Histogram Equalization for Contrast Enhancement*, vol. 1, pp. 69-73, IJCA, 2010.

(Shaw et Xu, [2009]) Shaw.G, Xu.Y, *Enhancing an Incremental Clustering Algorithm for Web Page Collections*, pp. 81-84, IJCWIIAT, 2009.

(Sherrir et Johnson, [1987]) Sherrir.H, Johnson.G, *Regionally adaptive histogram equalization of the chest*, vol. 6, pp. 1-7, MI, 1987.

(Siddiqi et al. [2002]) Siddiqi.K, Bouix.S, Tannenbaum.A, Zucker.S, *Hamilton-Jacobi skeletons*, vol. 3, pp. 215-231, IJCV, 2002.

(Siddiqi et Vincent, [2008]) Siddiqi.I, vincent.N, *How to Define Local Shape Descriptors for Writer Identification and Veri*fication, pp. 199-204, PRIS, 2008.

(Siddiqi et Vincent, [2009]) Siddiqi.I, vincent.N, *A set of chain code based features for writer recognition*, pp. 981–985, ICDAR, 2009.

(Siddiqi, [2009]) Siddiqi.I, *Classification of Handwritten Documents : Writer Recognition*, Thèse, 2009.

(Sivaraj et al. [2012]) Sivaraj.R, Ravichandran.T, Devi Priya.R, *Boosting Performance of Genetic Algorithm through Selective Initialization*, vol.68, pp.93-100, EJSR, 2012.

(Slagle et al. [1975]) Slagle.J, Chang.C, Heller.S, *A clustering and data-reorganizing algorithm*, vol. 5, pp. 125–128, SMC 1975.

(Spitz, [1990]) Spitz.A, *Multilingual Document Recognition*, pp. 193-206, ICEP, 1990.

(Spitz et al. [1994]) Spitz.A, Ozaki.M, *Palace: A Multilingual Document Recognition System*, pp. 16-37, IAPR, 1994.

(Su et al. [2009]) Su.Z, Cao.Z, Zhen.Y, *Identification of unreliable segments to improve skeletonization of handwriting images*, PAA, 2009.

(Sun, [1989]) Sun.Y, *Automated identification of arterial contours in coronay arteriograms by an adaptive tracking algorithm*, vol. 8, pp. 78-88, MI, 1989.

(Sung et al. [2006]) Sung.J, Bang.S, Choi.S, *A Bayesian network classifier and hierarchical Gabor features for Handwritten Numeral Recognition*, vol. 27, pp. 66-75, PRL, 2006.

(Tan, [1998]) Tan.T, *Rotation Invariant Texture Features and Their Use in Automatic Script Identification*, vol. 20, pp. 751-756, PAMI, 1998.

(Tan, [2002]) Tan.C, Cao.R, Shen.P, *Restoration of Archival Documents Using a Wavelet Technique*, pp. 1399-1404, PAMI, 2002.

(Tanaka, [2009]) Tanaka.H, *Threshold Correction of Document Image Binarization for Ruled-line Extraction*, pp. 541-545, ICDAR, 2009.

(Teague, [1980]) Teague.M, *Image analysis via the general theory of moments*, vol. 8, pp. 920-930, JOSA, 1980.

(Tho et Tang, [2001]) Tho.Y, Tang.Y, *Discrimination of Oriental and Euramerican Scripts Using Fractal Feature*, p. 1115-1119, ICDAR, 2001.

(Tjen-Sien, et al. [2000]) Tjen-Sien.L, Wei-Yin.L, Yu-Shan.S, *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms*, vol. 40, pp. 203– 228, ML, 2000.

(Tomai et al. [2004]) Tomai.C, Zhang.B, Srihari.S, *Discriminatory power of handwritten words for writer recognition*, vol. 2, pp. 638–641, ICPR, 2004.

(Trier et Jain, [1995]) Trier.O, Jain.A, *Goal-directed evaluation of binarization methods*, vol. 17, pp. 1191-1201, PAMI, 1995.

(Trier et Taxt, [1995]) Trier.O, Taxt.T, *Improvement of integrated function algorithm for binarization of document images*, vol. 16, pp. 277-283, PRL, 1995.

(Trincklin, [1984]) Trincklin.J, *Conception d'un système d'analyse de documents. Etude et réalisation d'un module d'extraction de la structure physique de document à support visuel*, 1984.

(Truong et Amini, [2008]) Truong.V, Amini.M, *Apprentissage de fonctions d'ordonnancement semi–supervisé inductives*, CAP, 2008.

(Tsai, [1985]) Tsai.W, *Moment-preserving thresholding: a new approach*, vol. 29, pp. 377-39, CVGIP, 1985.

(Tversky, [1975]) Tversky.A, *Features of similarity*, vol. 84, pp. 327-352, Psychological Review, 1977.

(Tversky, [1978]) Tversky.A, Gati.I, *Studies of similarity*, pp. 79–98,Cognition and categorization, 1978.

(Vapnic, [1995]) Vapnik.V, *The Nature of Statistical Learning Theory*, Springer, 1995.

(Vapnic, [1998]) Vapnik.V, *Statistical learning theory*. Wiley, 1998.

(Velasco, [1980]) Velasco.F, *Thresholding using the ISODATA clustering algorithm*, vol. 10, pp. 771-774, SMC, 1980.

(Vese et Chan, [2002]) Vese.L, Chan.T, *A multiphase level set framework for image segmentation using the Mumford and Shah model*, vol. 50, pp. 271–293, IJCV, 2002.

(Vincent et Frêche, [2001]) Vincent.N, Frêche.T, *Gray Level Use in a Handwriting Fractal Approach and Morphological Properties Quantification*, pp.307-311, ICDAR, 2001.

(Vishwakarma, [2012]) Vishwakarma.A, *Color Image Enhancement Techniques, A Critical Review*, vol. 3, pp. 39-45, IJCSE, 2012.

(Wake et al. [1998]) Waked.B, Bergler.S, Suen.C, Khoury.S, *Skew Detection, Page Segmentation and Script Classification of Printed Document Images*,vol. 5, pp. 4470-4475, ICSMC, 1998.

(Wall et Danielsson, [1984]) Wall.K, Danielsson.P, *A fast sequential method for polygonal approximation of digitized curves*, pp. 220–227, CVGIP, 1984.

(Wallace et Dowe, [1994]) Wallace.C, Dowe.D, *Intrinsic classification by mml-the snob program*, pp. 37-44, AJCAI, 1994.

(Wang et al. [2005]) Wang.J, Neskovic.P, Cooper.L, *A probabilistic model for cursive handwriting recognition using spatial context*, ICASSP, 2005.

(Wang et Pavlidis, [1993]) Wang.L, Pavlidis.T, *Direct gray-scale extraction of features for character recognition*, vol. 15, pp. 1053-1067, PAMI, 1993.

(Wang et Christofides, [2008]) Wang.X, Christofides.P, *Control of Particulate Processes*, vol. 25, pp. 287-387, PPSC, 2008.

(Wettschereck et al. [1997]) Wettschereck.D, Aha.D, Mohri.T, *A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms*, vol.11, pp. 273–314, AIR, 1997.

(Williams et Lambert, [1959]) Williams.W, J.Lambert.J, *Multivariate methods in plant ecology*, vol. 47, pp. 83-101, IAAPC, 1959.

(Wold et Doermann, [2002]) Wolf.C, Doermann.S, *Binarization of Low Quality Text Using a Markov Random Field Model*, vol. 3, pp. 160-163, ICPR, 2002.

(Wolf et al. [2002]) Wolf.C, Jolion.J, Chassaing.F, *Text Localization, Enhancement and binarization in Multimedia Documents*, vol. 4, pp. 1037-1040, ICPR, 2002.

(Wood et al. [1995]) Wood.S, Yao.X, Krishnamurthi.K, Dang.L, *Language Identification for Printed Text Independent of Segmentation*, vol. 3, pp. 428-431, ICIP, 1995.

(Wu et Amin, [2003]) Wu.S, Amin.A, *Automatic thresholding of gray-level using multistage approach*, vol. 1, pp. 493-497, ICDAR, 2003.

(Wu et Manmatha, [1998]) Wu.V, Manmatha.R, *Document image clean-up and binarization*, SPIE, 1998.

(Xu et al. [2007]) Xu.Y, Zhang.H, Li.H, Hu.G, *An improved algorithm for vessel centerline tracking in coronary angiograms*, vol. 88, pp. 131-143, CMPB, 2007.

(Yan et al. [2009]) Yan.Y, Chen.Q, Deng.W, Yuan.F, *Chinese Handwriting Identification Based on Stable Spectral Feature of Texture Images*, vol.2, IJIES, 2009.

(Yanowitz et Bruickstein, [1989]) Yanowitz.S, Bruickstein.A, *A new method for image segmentation*, vol. 46, pp. 82-95, CVGIP, 1989.

(Yao-Hon, [2007]) Yao-Hong.T, *A New Approach for Image Thresholding under Uneven Lighting Conditions*, pp. 123-127, ICIS, 2007.

(Yim et al. [2000]) Yim.P, Choyke.L, *Summers.R, Gray-Scale Skeletonization of Small Vessels in Magnetic Resonance Angiography*, vol. 19, MI, 2000.

(You et Tang, [2007]) You.X, Tang.Y, *Wavelet-based approach to character skeleton*, vol. 16, pp. 1220–1231, IP, 2007.

(Yu et Bajaj, [2004]) Yu.Z, Bajaj.C, *A Segmentation-Free Approach for Skeletonization of Gray-Scale Images via Anisotropic Vectore Diffusion*, pp. 1063-1069, CVPR, 2004.

(Zhu et Sun, [2002]) Zhang.H, Sun.G, *Feature selection using Tabu Search method*, vol.35, pp. 701-711, PR, 2002.

(Zhu et Suen, [1984]) Zhang.T, Suen.C, *A fast parallel algorithm for thinning digital patterns*, vol. 27, pp. 236–240, ACM, 1984.

(Zhang et al. [2009]) Zhang.Y, Zhou.X, Degterev.A, Lipinski.M, Adjeroh.D, Yuan.J, Wong.S, *A novel tracing algorithm for high throughout imaging Screening of neuron-based assays*, vol. 160, pp. 149-162, JNM, 2007.

(Zhu et al. [2009]) Zhu.G, Yu.X, Li.Y, Doermann.S, *Language identification for handwritten document images using a shape codebook*, vol.42, pp.3184-3191, PR, 2009.